Optimal Hospital Care Scheduling During the SARS-CoV-2 Pandemic

Josh C. D'Aeth*[1], Shubhechyya Ghosal*[2], Fiona Grimm*[3], David Haw*[1], Esma Koca*[2], Krystal Lau*[4], Huikang Liu*[5], Stefano Moret*[2], Dheeya Rizmie*[4], Peter C Smith[4], Giovanni Forchini†[1,6], Marisa Miraldo†[4], and Wolfram Wiesemann†‡[2]

[1]MRC Centre for Global Infectious Disease Analysis & WHO Collaborating Centre for Infectious Disease Modelling, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), School of Public Health, Imperial College London, London, UK.

[2]Department of Analytics, Marketing & Operations, Imperial College Business School, Imperial College London, London, UK.

[3]The Health Foundation London, UK.

[4]Department of Economics and Public Policy & Centre for Health Economics and Policy Innovation, Imperial College Business School, Imperial College London, London, UK.

[5]Research Institute for Interdisciplinary Sciences, School of Information Management and Engineering, Shanghai University of Finance and Economics.

[6]Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

Abstract

The COVID-19 pandemic has seen dramatic demand surges for hospital care that have placed a severe strain on health systems worldwide. As a result, policy makers are faced with the challenge of managing scarce hospital capacity so as to reduce the backlog of non-COVID patients whilst maintaining the ability to respond to any potential future increases in demand for COVID care. In this paper, we propose a nation-wide prioritization scheme that models each individual

*Contributed Equally

†Contributed Equally

‡Corresponding author: ww@imperial.ac.uk

patient as a dynamic program whose states encode the patient's health and treatment condition, whose actions describe the available treatment options, whose transition probabilities characterize the stochastic evolution of the patient's health and whose rewards encode the contribution to the overall objectives of the health system. The individual patients' dynamic programs are coupled through constraints on the available resources, such as hospital beds, doctors and nurses. We show that the overall problem can be modeled as a *grouped* weakly coupled dynamic program for which we determine near-optimal solutions through a fluid approximation. Our case study for the National Health Service in England shows how years of life can be gained by prioritizing specific disease types over COVID patients, such as injury & poisoning, diseases of the respiratory system, diseases of the circulatory system, diseases of the digestive system and cancer.

Keywords: COVID, Care Prioritization, Grouped Weakly Coupled Dynamic Programs, Fluid Approximation.

1 Introduction

Across health systems globally, hospitals struggle to meet the demand surges caused by the SARS-CoV-2 (hereafter COVID) pandemic. Despite the expansion of hospital capacity (*e.g.*, through field hospitals; McCabe et al. 2020 and Christen et al. 2021) and the implementation of lockdown measures to smooth over time the pressures on care provision, policy makers face an unprecedented challenge in managing scarce hospital capacity and treating non-COVID patients whilst maintaining the ability to respond to any potential future increases in demand for COVID care.

In this context, care prioritization policies become vital to mitigate the morbidity and mortality associated with surges of demand overflowing the existing capacity. Prioritization policies are common in health systems where the available resources are insufficient to cope with significant seasonal demand peaks (Rizmie et al., 2019). While those pressures are normally short-lived and do not require a drastic change in care prioritization or investments in extra capacity, the pressures of the COVID pandemic are more severe due to its prolonged duration, the uncertainty in the number of COVID patients that require care, the timing of the demand surges, the intensity of resource usage required to address the needs of COVID patients and the fact that the pandemic impacts the entire population. In response, several countries have deployed a wide range of prioritization policies to delay access to care for some patients who are perceived as requiring less urgent treatment

(NHS, 2020d; NICE, 2020). The National Health Service (NHS) in England, for example, provided national level guidance on the cancellation of non-urgent elective (*i.e.*, planned) procedures as well as a prioritization to intensive care of COVID patients below the age of 65 and with a high capacity to benefit (Gardner et al., 2020). The ethical guidelines published by the German Interdisciplinary Association for Intensive Care and Emergency Medicine discuss the prioritization to intensive care of COVID patients who do not suffer from severe respiratory illness (DIVI, 2020). The Italian College of Anesthesia, Analgesia, Resuscitation and Intensive Care advised the prioritization to intensive care of COVID patients above 70 years of age that do not have more than one admission per year for a range of diseases (Riccioni et al., 2020). We refer the reader to Joebges and Biller-Andorno (2020) for a review of the prioritization guidelines applied in different countries during 2020.

The aforementioned blanket policies tend to prioritize COVID patients in detriment to patients with other diseases without systematically accounting for the trade-offs between the provision of COVID and non-COVID care. For example, non-prioritized patients that see their planned care postponed or canceled might have a higher capacity to benefit from treatment than those prioritized; also, these patients' diseases might progress considerably while they wait for care, and they may subsequently require emergency or more complex treatment, thus creating further pressures on hospital capacity. As a result, blanket policies are likely to impact morbidity and mortality as well as increase the financial burden on health systems. Against this backdrop, the Nuffield Council on Bioethics, the UK's main health and healthcare ethics authority, has recently urged policy makers to develop optimal tools and national guidance to best allocate scarce hospital capacity to minimize the detrimental impact the pandemic has on population health (The Nuffield Council on Bioethics, 2020).

In this paper, we develop an optimization-based prioritization scheme that schedules patients into general & acute (G&A[1]) as well as critical care (CC[2]) so as to minimize overall years of life lost (YLL).[3] We consider a national-level scale (rather than an individual hospital) in order to inform strategic public health policy-making, which is particularly relevant in the case of a pandemic affecting an entire country. Our optimization scheme is dynamic and considers weekly patient cohorts subdivided into different patient groups (defined by disease, age group and admission type: elective

[1] G&A consists of beds allocated for curative care, such as relieving symptoms and managing illnesses or injuries, managing labor, performing surgeries, performing diagnostics and other medical procedures.

[2] CC refers to a specialized and separate area in a hospital, often called *intensive care unit* or *intensive therapy unit*, which is adequately staffed and equipped to manage and monitor patients with life-threatening conditions.

[3] YLL quantifies the years of life lost due to premature deaths, accounting for the age at which deaths occur.

and emergency) over a 56-weeks time horizon. We model each patient as a dynamic program (DP) whose states encode the patient's health status (proxied by the categorization elective/emergency, recovered or deceased) and treatment condition (waiting for treatment, in G&A or in CC), whose actions describe the treatment options (admit or move to G&A or to CC, deny care or discharge from hospital), whose transition probabilities characterize the stochastic evolution of the patient's health and whose rewards amount to the years of life gained. Our model simultaneously optimizes the treatment of all patients while accounting for capacity constraints on the supply side, including the availability of G&A as well as CC beds and staff (senior doctors, junior doctors and nurses). By clustering the patients into groups (defined through the same arrival time, disease type, age group and admission type) that can each be described through the same DP, we obtain a *grouped* weakly coupled DP that records for each patient group how many patients are in a particular state, and how many times which action is applied to those patients. We show that the classical fluid relaxation of ordinary weakly coupled DPs applies to grouped weakly coupled DPs as well and gives rise to a linear program (LP) that scales in the number of groups (as opposed to the number of constituent DPs). Moreover, the solutions to this LP allow us to recover near-optimal solutions to the grouped weakly coupled DP with high probability. We demonstrate the power of our modeling framework in a case study of the NHS in England, where we cluster approximately 10 million patients (the entire population in need of care) into 3,360 patient groups whose admission we manage over the course of one year in weekly granularity.

The contributions of this paper may be summarized as follows.

(i) We propose a class of weakly coupled DPs—the *grouped* weakly coupled DPs—which admit a fluid relaxation that scales gracefully with the problem data. The fluid relaxation allows us to recover provably high-quality solutions to the grouped weakly coupled DP.

(ii) We apply our findings to a case study of the NHS in England, where we show how grouped weakly coupled DPs allow to prioritize access to elective and emergency care. Our case study appears to be one of the largest and most detailed applications of weakly coupled DPs to date.

(iii) We publish the data of our case study as well as the source code of our prioritization scheme, so that it is available for researchers, practitioners and policy makers to develop further research and/or inform prioritization policies.

This paper is part of two related publications. The accompanying paper (D'Aeth et al., 2021) details our data collection and epidemiological modelling and uses the methodology developed in this paper to assess the effects of policies implemented by the English government to derive policy recommendations for the NHS in England. In contrast, the present paper develops the theoretical foundations of the applied analysis by D'Aeth et al. (2021): we develop the concept of grouped weakly coupled DPs, we propose their approximation via fluid DPs and the subsequent solution via LPs, and we show how to recover provably high-quality solutions to the original grouped weakly coupled DP from solutions to these LPs.

The methodology and application of our paper builds upon a rich body of medical and methodological literature, which we review in the remainder of this section.

Under normal operation, elective care is typically scheduled via prioritization schemes (MacCormick et al., 2003; Déry et al., 2020). Recent months have seen a rapidly growing body of literature that discusses the scheduling of elective care surgeries in light of the COVID pandemic. In contrast to our work, which studies hospital care in a broader sense, the majority of that literature focuses on prioritization of COVID care (*e.g.*, Phua et al. 2020) or surgeries (mostly for cancer), with most papers evaluating the impact of the pandemic on elective surgeries (Fujita et al., 2020; Negopdiev et al., 2020; Sud et al., 2020; Yoon et al., 2020), proposing guidelines based on best practices in individual hospitals (Argenziano et al., 2020; Eichberg et al., 2020; Tzeng et al., 2020) or reviewing the guidelines of national authorities (Burki, 2020). These guidelines are developed by domain experts and tend to be qualitative in nature (Moris and Felekouras, 2020; Soltany et al., 2020). In contrast, Bertsimas, Lukin et al. (2020), Bertsimas, Pauphilet et al. (2020), Davis et al. (2020), Gao et al. (2020) and Vaid et al. (2020) employ machine learning techniques (such as support vector machines, tree ensembles and neural networks) to estimate the mortality risk of COVID patients, which can subsequently be used as a proxy of need for patient prioritization. While these contributions are important, they highlight prioritization schemes within a specific disease or subgroup of patients or care settings within a single hospital, and they are static and thus consider neither the dynamic nature of surges in demand nor the complexity of the dynamic needs of patients. In contrast, we propose a national prioritization scheme across all disease groups that accounts for future demand surges and capacity fluctuations as well as the evolution of the patients' needs over the course of the pandemic, which to the best of our knowledge has not been proposed so far.

As an alternative to static prioritization schemes, the operations research literature has studied the dynamic management of G&A and CC capacity via admissions and discharge policies. For example, Bekker and Koeleman (2011) and Meng et al. (2015) propose capacity management policies for G&A beds using queueing theory and robust optimization, Chan et al. (2012), Kim et al. (2015) and Ouyang et al. (2020) develop capacity management policies for CC beds via queueing theory, DPs and simulation, and Helm et al. (2011) and Shi et al. (2019) study hospital-wide capacity management policies using DPs. These approaches aim to optimize the often conflicting goals of short-term and long-term patient welfare as well as hospital costs, subject to constraints on the available resources. In contrast to our work, these papers focus on individual hospitals, which allows them to model hospital operations and within-hospital patients' care pathways at a finer granularity: admissions and discharge decisions are often taken at an hourly granularity, and some models account for longer-term implications of decisions such as readmissions.

From a methodological viewpoint, our work contributes to the rich theory of weakly coupled DPs, which aim to alleviate the *curse of dimensionality* that plagues classical DPs (Bertsekas, 1995; Puterman, 2014) by decomposing them into independently evolving constituent DPs that are coupled by a small number of resource constraints. The decomposition of generic weakly coupled DPs appears to have been pioneered by Hawkins (2003). The author shows that by dualizing the coupling constraints, a weakly coupled DP decomposes into set of constituent DPs whose optimal cost to-go functions can be determined efficiently. This insight is used to determine optimal Lagrange multipliers and cost to-go functions for the constituent DPs by an LP or a stochastic subgradient method, and to derive feasible (but suboptimal) policies for the weakly coupled DP via three different heuristics. Adelman and Mersereau (2008) present the problem dual to the LP relaxation proposed by Hawkins (2003), which in our context gives rise to what we term the 'fluid relaxation'. They also compare the Lagrangian relaxation of Hawkins (2003) with an LP-based relaxation of the weakly coupled DP that imposes additively separable value functions for all constituent DPs, and they show that the LP-based relaxation, while substantially more challenging to solve due to its exponential number of constraints, provides weakly better bounds than the Lagrangian relaxation. Bertsimas and Mišić (2016) build upon and extend the work of Hawkins (2003) and Adelman and Mersereau (2008) by studying weakly coupled DPs with generic action spaces that do not necessarily result from cross products of the action spaces of the constituent DPs intersected with resource constraints.

They show that the Lagrangian and the LP-based relaxations apply to their model as well, and they refine these relaxations by considering partially time-disaggregated formulations as well as higher-order formulations that account for the stochastic dependencies between the constituent DPs.

The aforementioned approaches have in common that they do not offer any bounds on the suboptimality incurred by the derived policies, except for their optimality in very specific settings that are mainly of theoretical interest. Indeed, most performance guarantees have been developed for specific classes of weakly coupled DPs. Caro and Gallien (2007) apply the Lagrangian relaxation principle to a finite horizon multi-armed bandit problem in which the decision maker can pull multiple arms in each time period. They use their approach to model a dynamic assortment optimization problem faced by fast fashion companies, who need to decide how to update their assortments throughout the sales season based on the observed demands. Their case study appears to be one of the largest and most detailed real-world applications of weakly coupled DPs to date. Hu and Frazier (2017) and Zayas-Cabán et al. (2019) apply the Lagrangian relaxation principle to restless bandit problems where each bandit is governed by the same transition kernel, and they prove asymptotic optimality of the resulting policies as the number of arms approaches infinity. Brown and Smith (2020) study the dynamic selection problem, which is a generalization of the finite horizon restless bandit problem that allows for non-stationary dynamics, and use Lagrangian relaxation to derive a policy whose reward falls short of the optimal one by at most $\mathcal{O}([\overline{r} - \underline{r}] \cdot 2^T \cdot \sqrt{N})$, where $\overline{r}$ and $\underline{r}$ represent upper and lower bounds on the reward achievable by any arm in any time stage, T is the time horizon and N is the maximum number of arms that can be pulled in any time period. Under mild assumptions, this bound implies asymptotic optimality of the derived policy when the number of arms approaches infinity. Marklund and Rosling (2012) and Nambiar et al. (2021) study a supply chain stock allocation problem where an existing inventory is distributed to a number of retailers over a finite time horizon so as to serve stochastic demands. The authors develop a relaxation that aggregates the inventory constraints and enforces feasibility in expectation; this is reminiscent of the Lagrangian decomposition of weakly coupled DPs, which can be shown to enforce the coupling resource constraints in expectation. The authors dualize the remaining expected inventory constraint and thereby decouple the problem into individual problems for each retailer. Their approaches result in lower bounds and feasible policies which incur a suboptimality of $\mathcal{O}(T^{5/2} \cdot \sqrt{N})$ and $\mathcal{O}(T \cdot \sqrt{N})$, respectively, where N is the number of retailers and T is the time horizon; in particular, both the lower bounds and the feasible policies are asymptotically optimal as the number of retailers approaches infinity. These results have subsequently been strengthened by Miao et al. (2021) to a suboptimality of $\mathcal{O}(\sqrt{TN \cdot \log TN})$, who also extend the analysis to a multi-warehouse multi-retailer setting.

To our best knowledge, Brown and Zhang (2020) contribute the only prior work that offers suboptimality bounds for generic weakly coupled DPs in the sense of Hawkins (2003). In fact, the authors study a generalization of weakly coupled DPs where an observable signal, which is governed by an exogenous finite-state Markov process, affects the transitions, rewards and/or resource consumption of the constituent DPs. The authors decompose the weakly coupled DP via Lagrangian relaxation and develop an iterative primal-dual algorithm to solve the emerging large-scale fluid LP. They derive a feasible policy whose suboptimality, in the absence of the shared signal, is bounded by $\mathcal{O}([\overline{r} - \underline{r}] \cdot 2^T \cdot \sqrt{N})$, where $\overline{r}$ and $\underline{r}$ represent upper and lower bounds on the rewards achievable by any constituent DP in any time stage, N is the number of constituent DPs and T is the time horizon; in particular, the policy is asymptotically optimal as N approaches infinity.

Our main methodological contribution is the introduction of *grouped* weakly coupled DPs, which are weakly coupled DPs whose constituent DPs form groups that share the same transition and reward structure. The existence of groups allows us to solve the fluid relaxations of weakly coupled DPs with millions of constituent DPs, which are beyond the reach of the existing solution approaches. Moreover, our fluid relaxation allows us to recover feasible solutions. As the number N of patients increases, the suboptimality of our feasible policy is bounded by $\mathcal{O}(T \cdot \sqrt{N \log N} + (|\mathcal{T}|^2 \cdot |\mathcal{L}|)/N)$ in expectation as well as—in any particular run of the grouped weakly coupled DP—by $\mathcal{O}(T \cdot \sqrt{N \log N})$ with high probability; here, T denotes the time horizon. These bounds compare favourably with the existing bounds, and to our best knowledge, similar high-probability bounds for any particular run of a weakly coupled DP have not been discussed in the existing literature.

While the concept of grouped weakly coupled DPs was developed with the outlined healthcare application in mind, we emphasize its applicability in other applications as well, such as B2C marketing where current and prospective customers should be assigned to marketing campaigns based on their purchase likelihood. Here, customers can be modelled as DPs whose states encode the current product portfolio as well as preferences learnt from previous campaigns, and whose actions describe the inclusion to (or exclusion from) a particular campaign.

Beyond the context of weakly coupled DPs, fluid relaxations have been used in different application domains to study asymptotic settings where the systems under consideration grow in size. Maglaras (2000), for example, studies the fluid relaxation of a multi-class queueing network and derives a policy from the relaxation that is asymptotically optimal as the backlog and the system observation time increase. Bertsimas et al. (2003) consider the fluid relaxation of a high-multiplicity job-shop scheduling problem and derive a policy from the relaxation that is asymptotically optimal as the numbers of jobs from each type increase. Armony and Maglaras (2004) study the optimal operation of contact centers with a real-time telephone service and a postponed call-back service. Using the concept of fluid scaled processes, they develop policies that are asymptotically optimal as the number of agents grows to infinity and the traffic intensity grows to one. Tsitsiklis and Xu (2012) study a multi-server queueing system in which the available resources can be split between centralized and distributed employment. The authors use a fluid relaxation to analyze the benefits of resource pooling, and they show that the behavior of the actual system converges to that of the fluid approximation as the available resources and the arrival rate of the requests increase. Maglaras and Meissner (2006) consider fluid relaxations of multi-product revenue management problems and propose policies for the associated pricing and capacity control problems that are asymptotically optimal as the potential demand and capacity increase. Fluid approximations of counting models have also been studied in performance analysis, where they are known as mean-field approximations (Benaïm and Le Boudec, 2008), and in the analysis of biochemical networks (Ciocchetta et al., 2009).

The remainder of the paper proceeds as follows. Section 2 introduces grouped weakly coupled DPs, which will serve as our model of the health system. Section 3 shows that grouped weakly coupled DPs are amenable to a similar fluid approximation as non-grouped weakly coupled DPs, and that this fluid approximation allows us to obtain high-quality solutions to the grouped weakly coupled DP in polynomial time. Section 5 discusses our case study of the NHS in England. Section 6 concludes with a discussion of the limitations of our healthcare model. For ease of exposition, all proofs are relegated to the appendix.

Notation. For a finite set $\mathcal{X} = \{1, \ldots, X\}$, we denote by $\Delta(\mathcal{X})$ the set of all probability distributions supported on $\mathcal{X}$, that is, all functions $p : \mathcal{X} \to \mathbb{R}_+$ satisfying $\sum_{x \in \mathcal{X}} p(x) = 1$. For a logical expression $\mathcal{E}$, we let $\mathbf{1}[\mathcal{E}] = 1$ if $\mathcal{E}$ is true and $\mathbf{1}[\mathcal{E}] = 0$ otherwise.

2 Grouped Weakly Coupled Dynamic Programs

We first review how multiple DPs, each of which competes for the same set of resources, can be aggregated to a weakly coupled DP. We subsequently introduce the notion of a grouped weakly coupled DP, which stipulates that some of the DPs within a weakly coupled DP share the same characteristics. Section 3 will develop a fluid approximation for grouped weakly coupled DPs whose tightness is directly related to the sizes of these groups. Our healthcare application in Section 5, finally, will model individual patients as DPs, and patients with similar characteristics (arrival time, age group and disease type) will naturally form groups.

We model individual patients, which form the basis of our healthcare model, as DPs whose states record the patient's health (elective/emergency, recovered or deceased) and treatment state (waiting for treatment, in G&A or in CC), whose actions describe the treatment options (admit or move to G&A or to CC, deny care or discharge from hospital), whose transition probabilities characterize the stochastic evolution of the patient's health and whose rewards amount to the years of life gained.

Definition 1 (DP)**.** *For a finite time horizon $\mathcal{T} = \{1, \ldots, T\}$, a* DP *is specified by the tuple $(\mathcal{S}, \mathcal{A}, q, p, r)$, where $\mathcal{S} = \{1, \ldots, S\}$ denotes the finite state space, $\mathcal{A} = \{1, \ldots, A\}$ is the finite action space with $\mathcal{A}_t(s) \subseteq \mathcal{A}$ the admissible actions in state $s \in \mathcal{S}$ at time $t \in \mathcal{T}$, $q \in \Delta(\mathcal{S})$ are the initial state probabilities, $p = \{p_t\}_t$ with $p_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, $t \in \mathcal{T}$, are the Markovian transition probabilities, and $r = \{r_t\}_t$ with $r_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $t \in \mathcal{T}$, are the expected rewards.*

In a DP, a *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathcal{S} \to \mathcal{A}$ specifies for each time period $t \in \mathcal{T}$ and each state $s \in \mathcal{S}$ what action $\pi_t(s) \in \mathcal{A}$ is taken. A feasible policy π must satisfy $\pi_t(s) \in \mathcal{A}_t(s)$ for all $t \in \mathcal{T}$ and $s \in \mathcal{S}$. Under the policy π, a DP evolves as follows. The initial state $\tilde{s}_1$ is random and satisfies $\mathbb{P}[\tilde{s}_1 = s] = q(s)$ for $s \in \mathcal{S}$. For $t \in \mathcal{T} \backslash \{T\}$, the transitions are governed by

$$\mathbb{P}\left[\tilde{s}_{t+1} = s'\right] \;=\; \sum_{s \in \mathcal{S}} p_t(s' \,|\, s, \pi_t(s)) \cdot \mathbb{P}\left[\tilde{s}_t = s\right] \qquad \forall s' \in \mathcal{S}.$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t \in \mathcal{T}} r_t(\tilde{s}_t, \pi_t(\tilde{s}_t))\right]$.

Example 1 (DP)**.** *Figure 1 illustrates a DP with the states* 1 *and* 2 *and the actions A (admissible in both states) and B (admissible in state* 2 *only). Under action A, the system transitions to either*

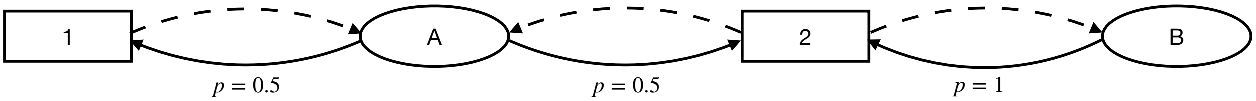

Figure 1. DP with two states and two actions. The rectangular and oval nodes represent the states and actions, respectively. Each dashed line represents the choice of an action in a period t, whereas the solid lines represent the state transitions from period t to period $t+1$.

state with probability 1/2, whereas the system remains in state 2 if action B is taken. The expected rewards are $r_t(1, \mathrm{A}) = 0$ and $r_t(2, \mathrm{A}) = r_t(2, \mathrm{B}) = 1$. As a result, the unique optimal policy π takes action A in state 1 and action B in state 2, respectively.

Our healthcare model combines all individual patient DPs to a single DP that records the state of each patient while also restricting the admissible policies to those that satisfy certain resource constraints (*e.g.*, the availability of G&A and CC beds, nurses and doctors).

Definition 2 (Weakly Coupled DP)**.** *For a finite set of DPs $(\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i)$, $i \in \mathcal{I} = \{1, \ldots, N\}$, over the same time horizon $\mathcal{T} = \{1, \ldots, T\}$, the* weakly coupled DP $(\{\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i\}_i)$ *is the DP $(\mathcal{S}, \mathcal{A}, q, p, r)$ with state space $\mathcal{S} = \times_{i \in \mathcal{I}} \mathcal{S}_i$, action space $\mathcal{A} = \times_{i \in \mathcal{I}} \mathcal{A}_i$ with $\mathcal{A}_t(s) = \times_{i \in \mathcal{I}} \mathcal{A}_{it}(s_i)$, $s \in \mathcal{S}$, and*

$$\mathcal{A}_t^{\mathrm{C}}(s) \;=\; \left\{ a \in \mathcal{A}_t(s) \;:\; \sum_{i \in \mathcal{I}} c_{tli}(s_i, a_i) \leqslant b_{tl} \;\; \forall l \in \mathcal{L} \right\},$$

initial state probabilities $q(s) = \prod_{i \in \mathcal{I}} q_i(s_i)$ for $s \in \mathcal{S}$, transition probabilities $p_t(s' \mid s, a) = \prod_{i \in \mathcal{I}} p_{it}(s'_i \mid s_i, a_i)$ for $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$ and expected rewards $r_t(s, a) = \sum_{i \in \mathcal{I}} r_{it}(s_i, a_i)$.

Weakly coupled DPs have been studied, among others, by Hawkins (2003) and Adelman and Mersereau (2008). In a weakly coupled DP, the admissible actions $a \in \mathcal{A}_t^{\mathrm{C}}(s)$ in state $s \in \mathcal{S}$ must satisfy the constraints $a_i \in \mathcal{A}_{it}(s_i)$ of the individual DPs $i \in \mathcal{I}$ as well as the coupling resource constraints $a \in \mathcal{A}_t^{\mathrm{C}}(s)$. In particular, the feasibility of an action $a_i \in \mathcal{A}_i$ for the i-th constituent DP is not just determined by the state $s_i \in \mathcal{S}_i$, but it depends (through the resource constraints) on the states $s_{i'} \in \mathcal{S}_{i'}$, $i' \in \mathcal{I} \backslash \{i\}$, of the other constituent DPs as well. The constraints in $\mathcal{A}_t^{\mathrm{C}}$ allow us to model the resource consumption of individual patients (such as a G&A or CC bed, as well as fractions of doctor and nurse times – each of which can be modeled as a distinct resource $l \in \mathcal{L}$). Throughout the paper, we assume that $c_{tli} \geqslant 0$ as well as $b_{tl} > 0$ for all $t \in \mathcal{T}$, $l \in \mathcal{L}$ and $i \in \mathcal{I}$.

In a weakly coupled DP, a *policy* $\pi = \{\pi_t\}_t$ with $\pi_t : \mathcal{S} \to \mathcal{A}$ specifies for each time period $t \in \mathcal{T}$,

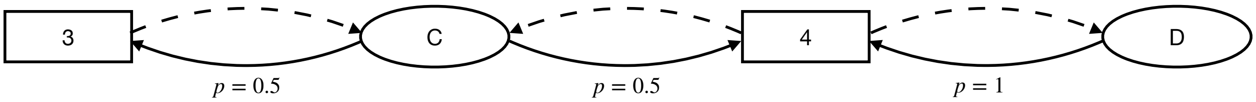

Figure 2. DP with two states 3 and 4 as well as two actions C (admissible in both states) and D (admissible in state 4 only). The expected rewards are $r_t(3, \mathrm{C}) = 0$ and $r_t(4, \mathrm{C}) = r_t(4, \mathrm{D}) = 1$.

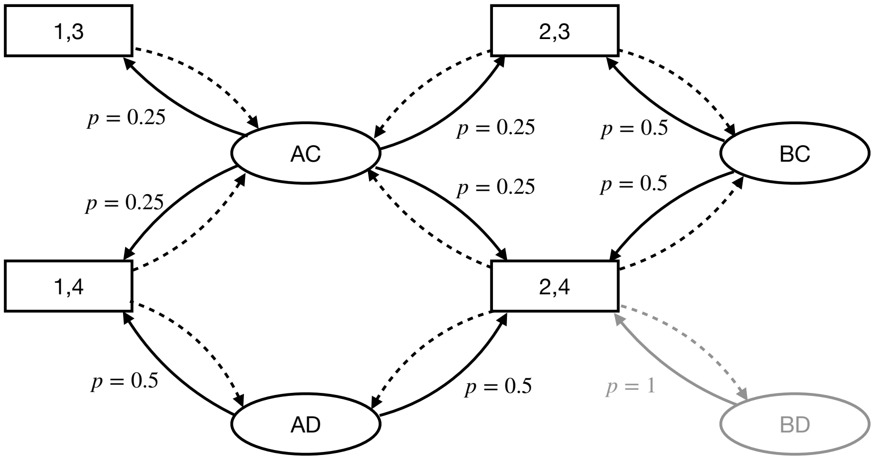

Figure 3. A weakly coupled DP is itself a DP. The values in the rectangular nodes record the states of the first and second DP, while the letters in the oval nodes denote the actions applied to each DP. The node corresponding to the action BD is shown in grey as this action violates the resource constraint $\mathbf{1}[s_1 = 2 \wedge a_1 = \mathrm{B}] + \mathbf{1}[s_2 = 4 \wedge a_2 = \mathrm{D}] \leqslant 1$ that requires at most one of the actions B and D to be selected at any time.

each DP $i \in \mathcal{I}$ and each state $s \in \mathcal{S}$ what action $[\pi_t(s)]_i \in \mathcal{A}_i$ is selected. A feasible policy π must satisfy $\pi_t(s) \in \mathcal{A}_t^{\mathrm{C}}(s)$ for all $t \in \mathcal{T}$ and $s \in \mathcal{S}$. We emphasize that the policy can choose the action $[\pi_t(s)]_i \in \mathcal{A}_i$ for the i-th DP in view of the states of all other constituent DPs, rather than just the state s_i; this is important in view of satisfying the coupling constraints. Under π, a weakly coupled DP evolves as follows. The initial state $\tilde{s}_1$ is random and satisfies $\mathbb{P}[\tilde{s}_1 = s] = \prod_{i \in \mathcal{I}} q_i(s_i) = q(s)$ for $s \in \mathcal{S}$. For $t \in \mathcal{T} \backslash \{T\}$, the transitions are governed by

$$\mathbb{P}\left[\tilde{s}_{t+1} = s'\right] = \sum_{s \in \mathcal{S}} \left[\prod_{i \in \mathcal{I}} p_{it}(s'_i \,|\, s_i, [\pi_t(s)]_i) \right] \cdot \mathbb{P}\left[\tilde{s}_t = s\right] = \sum_{s \in \mathcal{S}} p_t(s' \,|\, s, \pi_t(s)) \cdot \mathbb{P}\left[\tilde{s}_t = s\right] \qquad \forall s' \in \mathcal{S}.$$

The expected total reward of a policy π is $\mathbb{E}\left[\sum_{t \in \mathcal{T}} r_t(\tilde{s}_t, \pi_t(\tilde{s}_t))\right]$.

Example 2 (Weakly Coupled DP)**.** *Figure 3 combines the DP from Example 1 with the DP from Figure 2 to a weakly coupled DP that is subjected to the resource constraint* $\mathbf{1}[s_1 = 2 \wedge a_1 =$

B$] + \mathbf{1}[s_2 = 4 \wedge a_2 = D] \leqslant 1$*, that is, at most one of the actions B and D can be selected in any period. As a result, every optimal policy* π *selects one of the actions B or D (but not both) whenever they are admissible.*

Since it offers a lossless aggregation of its constituent DPs, the state and action spaces of a weakly coupled DP exhibit an undesirable scaling behavior:

$$|\mathcal{S}| = \prod_{i \in \mathcal{I}} |\mathcal{S}_i| \quad \text{and} \quad |\mathcal{A}| = \prod_{i \in \mathcal{I}} |\mathcal{A}_i|$$

The health system that we intend to model contains approximately 10 million patient DPs with 15 possible states and 6 possible actions each, thus resulting in a weakly coupled DP with approximately $15^{10{,}000{,}000}$ states and $6^{10{,}000{,}000}$ actions. Clearly, such a weakly coupled DP is not by itself amenable to an exact solution via one of the standard approaches such as value or policy iteration. In fact, even its approximate solution via Lagrangian relaxation (Hawkins, 2003; Adelman and Mersereau, 2008) remains challenging since the numbers of variables and constraints in the associated fluid relaxation scale linearly in the number of constituent DPs.

Our proposed fluid approximation for weakly coupled DPs relies on the following assumption, which stipulates that many of the DPs in the weakly coupled DP share the same dynamics.

Assumption 1 (Grouped Weakly Coupled DP)**.** *For the weakly coupled DP of Definition 2, we assume that there are* $J \ll N$ *groups such that each DP* $i \in \mathcal{I}$ *belongs to exactly one of the groups* $j \in \mathcal{J} = \{1, \ldots, J\}$*, and any two DPs of the same group share the same state and action spaces, initial state and transition probabilities as well as expected rewards.*

Assumption 1 implies that there exists a bijection between the set $\mathcal{I}$ of all DPs in the weakly coupled DP and the pairs $(j, i) \in \mathcal{J} \times \{1, \ldots, n_j\}$ that index the DPs $i \in \{1, \ldots, n_j\}$ belonging to the same group $j \in \mathcal{J}$. This bijection is unique up to reorderings of the groups as well as the DPs within each group. In the following, we fix any such bijection and refer to individual DPs of the grouped weakly coupled DP via their index pairs (j, i), and we denote the common description of the DPs in group $j \in \mathcal{J}$ by $(\mathcal{S}_j, \mathcal{A}_j, q_j, p_j, r_j)$. Our healthcare application will comprise approximately 10 million patients that form 3,360 patient groups characterized by the same arrival time $t \in \mathcal{T}$ as well as the same age group and disease type. It will then be natural to assume that the patients of the same patient group follow similar dynamics, as specified by the initial state and transition probabilities, and lead to similar conditional health outcomes, as specified by the expected rewards.

3 Fluid Approximation

Grouped weakly coupled DPs lend themselves to a linear programming approximation which is particularly suitable when the number N of DPs is large relative to the number J of groups, as is the case in our healthcare application. In the following, Section 3.1 develops and motivates this approximation, and Section 3.2 shows that the approximation allows us to recover high-quality solutions to the weakly coupled DP with high probability.

3.1 Linear Programming Formulation

Fix a weakly coupled DP $(\{\mathcal{S}_i, \mathcal{A}_i, q_i, p_i, r_i\}_i)$ satisfying Assumption 1. In any time period $t \in \mathcal{T}$ and state $s_t \in \mathcal{S}$, the feasible policies $\pi = \{\pi_\tau\}_{\tau=t}^T$ that maximize the expected total reward over the remaining time horizon $\tau \in \{t, \ldots, T\}$ are precisely the maximizers of the value equation

$$\begin{aligned} V_t(s_t) \;\; = \;\; \max_\pi \Bigg\{ & \mathbb{E}\left[\sum_{\tau=t}^{T} \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} r_{j\tau}(\tilde{s}^\pi_{\tau,(j,i)}, \tilde{a}^\pi_{\tau,(j,i)}) \;\middle|\; \tilde{s}^\pi_t = s_t \right] : \\ & \pi_\tau : \mathcal{S} \to \mathcal{A} \text{ with } \pi_\tau(s) \in \mathcal{A}^{\mathrm{C}}_\tau(s) \;\; \forall s \in \mathcal{S}, \; \forall \tau = t, \ldots, T \Bigg\}, \end{aligned}$$

where $\tilde{s}^\pi_{\tau,(j,i)}$ and $\tilde{a}^\pi_{\tau,(j,i)}$ denote the random state and action of the i-th DP in group j at time τ under the policy π, respectively. We next consider a Lagrangian relaxation that penalizes violations of the resource constraints with a non-negative penalty $\lambda = \{\lambda_{\tau l}\}_{\tau,l}$:

$$\begin{aligned} V^\lambda_t(s_t) \;\; = \;\; \max_\pi \Bigg\{ & \mathbb{E}\Bigg[\sum_{\tau=t}^{T} \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} r_{j\tau}(\tilde{s}^\pi_{\tau,(j,i)}, \tilde{a}^\pi_{\tau,(j,i)}) + \\ & \sum_{\tau=t}^{T} \sum_{l \in \mathcal{L}} \lambda_{\tau l} \left(b_{\tau l} - \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} c_{\tau l j}(\tilde{s}^\pi_{\tau,(j,i)}, \tilde{a}^\pi_{\tau,(j,i)}) \right) \;\middle|\; \tilde{s}^\pi_t = s_t \Bigg] : \qquad (1) \\ & \pi_\tau : \mathcal{S} \to \mathcal{A} \text{ with } \pi_\tau(s) \in \mathcal{A}_\tau(s) \;\; \forall s \in \mathcal{S}, \; \forall \tau = t, \ldots, T \Bigg\}, \end{aligned}$$

where we denote by $c_{\tau l j}$, $\tau \in \mathcal{T}$, $l \in \mathcal{L}$ and $j \in \mathcal{J}$, the common resource consumption function $c_{\tau l i}$ shared by all DPs (j, i), $i = 1, \ldots, n_j$, of group j. In contrast to the previous value equation, (1) only requires the policy π to satisfy the constraints $[\pi_\tau(s)]_i \in \mathcal{A}_{i\tau}(s_i)$ for the individual DPs $i \in \mathcal{I}$, whereas violations of the coupling resource constraints are tolerated (but penalized).

The following statement is well-known and summarized for convenience only.

Proposition 1. *The Lagrange value function* (1) *satisfies the following.*

(i) **Weak Duality.** *We have $V_t^\lambda(s_t) \geqslant V_t(s_t)$ for all $t \in \mathcal{T}$ and $s_t \in \mathcal{S}$.*

(ii) **Decomposition.** *We have*

$$V_t^\lambda(s_t) \quad = \quad \sum_{\tau=t}^{T} \sum_{l \in \mathcal{L}} \lambda_{\tau l} b_{\tau l} + \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} V_{t,(j,i)}^\lambda(s_{t,(j,i)}) \qquad \forall t \in \mathcal{T},\ \forall s_t \in \mathcal{S}, \tag{2}$$

where the DP-wise value functions $V_{t,(j,i)}^\lambda$, $(j,i) \in \mathcal{J} \times \{1,\ldots,n_j\}$ and $t \in \mathcal{T}$, satisfy

$$V_{t,(j,i)}^\lambda(s) \quad = \quad \max_{a \in \mathcal{A}_{jt}(s)} \left\{ r_{jt}(s,a) - \sum_{l \in \mathcal{L}} \lambda_{tl} c_{tlj}(s,a) + \sum_{s' \in \mathcal{S}_j} p_{jt}(s' \mid s,a) \cdot V_{t+1,(j,i)}^\lambda(s') \right\} \quad \forall s \in \mathcal{S}_j$$

and $V_{T+1,(j,i)}^\lambda(s) = 0$, $s \in \mathcal{S}_j$.

The first property of Proposition 1 follows immediately from the fact the Lagrangian value equation results from a relaxation of the original value equation. The second property shows that by relaxing the coupling resource constraints, the Lagrangian value function decomposes into a sum of individual value functions for each constituent DP $(j,i) \in \mathcal{J} \times \{1,\ldots,n_j\}$. Note that in our case, all DP-wise value functions $V_{t,(j,i)}^\lambda(s)$, $i \in \{1,\ldots,n_j\}$, of the same group $j \in \mathcal{J}$ are identical by construction, and they can thus be replaced by a single function V_{tj}^λ satisfying

$$V_{tj}^\lambda(s) \quad = \quad \max_{a \in \mathcal{A}_{jt}(s)} \left\{ r_{jt}(s,a) - \sum_{l \in \mathcal{L}} \lambda_{tl} c_{tlj}(s,a) + \sum_{s' \in \mathcal{S}_j} p_{jt}(s' \mid s,a) \cdot V_{t+1,j}^\lambda(s') \right\} \qquad \forall s \in \mathcal{S}_j,$$

$t \in \mathcal{T}$, with $V_{T+1,j}^\lambda(s) = 0$, $s \in \mathcal{S}_j$. Following standard results from the literature on weakly coupled DPs (see, *e.g.*, p. 5 of Adelman and Mersereau, 2008), we can then formulate the problem that maximizes the Lagrangian relaxation over all non-negative penalties as the following LP:

$$
\begin{array}{lll}
\underset{\lambda,\, V^{\lambda}}{\text{minimize}} & \displaystyle\sum_{t\in\mathcal{T}}\sum_{l\in\mathcal{L}}\lambda_{tl}b_{tl} + \sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j} n_j\cdot q_j(s)\cdot V^{\lambda}_{1j}(s) & \\
\text{subject to} & \displaystyle V^{\lambda}_{tj}(s) \geqslant r_{jt}(s,a) - \sum_{l\in\mathcal{L}}\lambda_{tl}c_{tlj}(s,a) + \sum_{s'\in\mathcal{S}_j} p_{jt}(s' \mid s,a)\cdot V^{\lambda}_{t+1,j}(s') & \forall t\in\mathcal{T},\ \forall j\in\mathcal{J}, \\
 & & \forall s\in\mathcal{S}_j,\ \forall a\in\mathcal{A}_{jt}(s) \\
 & \lambda_{tl}\geqslant 0 & \forall t\in\mathcal{T},\ \forall l\in\mathcal{L} \\
 & V^{\lambda}_{T+1,j}(s) = 0 & \forall j\in\mathcal{J},\ \forall s\in\mathcal{S}_j
\end{array}
\tag{3}
$$

We emphasize that the numbers of decision variables and constraints in this LP scale with the number of time periods T, the number of DP groups J, the number of resources L as well as the sizes of the individual state and action spaces, but they do *not* depend on the number N of constituent DPs anymore. This is an immediate consequence of Assumption 1 and Proposition 1, and it will be fundamental to the tractability of our approach.

For the subsequent analysis it is more convenient to work with the problem dual to the Lagrangian relaxation (3).

Proposition 2 (Fluid LP). *The strong linear programming dual of problem* (3) *is*

$$
\begin{array}{lll}
\underset{\sigma,\,\pi}{\text{maximize}} & \displaystyle\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\pi_{tj}(s,a) & \\
\text{subject to} & \sigma_{1j}(s) = n_j\cdot q_j(s) & \forall j\in\mathcal{J},\ \forall s\in\mathcal{S}_j \\
 & \displaystyle\sigma_{t+1,j}(s') = \sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} p_{jt}(s'\,|\,s,a)\cdot\pi_{tj}(s,a) & \forall j\in\mathcal{J},\ \forall s'\in\mathcal{S}_j,\ \forall t\in\mathcal{T}\backslash\{T\} \\
 & \displaystyle\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\pi_{tj}(s,a)\leqslant b_{tl} & \forall l\in\mathcal{L},\ \forall t\in\mathcal{T} \\
 & \displaystyle\sum_{a\in\mathcal{A}_j}\pi_{tj}(s,a) = \sigma_{tj}(s) & \forall j\in\mathcal{J},\ \forall s\in\mathcal{S}_j,\ \forall t\in\mathcal{T} \\
 & \pi_{tj}(s,a) = 0 & \forall j\in\mathcal{J},\ \forall s\in\mathcal{S}_j,\ \forall a\in\mathcal{A}_j\backslash\mathcal{A}_{jt}(s),\ \forall t\in\mathcal{T} \\
 & \sigma_{tj}(s),\ \pi_{tj}(s,a)\geqslant 0 & \forall j\in\mathcal{J},\ \forall (s,a)\in\mathcal{S}_j\times\mathcal{A}_j,\ \forall t\in\mathcal{T}.
\end{array}
\tag{4}
$$

The dual LP comprises $\mathcal{O}\left(|\mathcal{T}|\cdot\sum_{j\in\mathcal{J}}|\mathcal{S}_j||\mathcal{A}_j|\right)$ decision variables and $\mathcal{O}\left(|\mathcal{T}|\cdot\max\left\{\sum_{j\in\mathcal{J}}|\mathcal{S}_j|,\ |\mathcal{L}|\right\}\right)$ constraints. Our healthcare model comprises $|\mathcal{T}| = 56$ time periods (one year in weekly granularity, plus 4 weeks to alleviate end-of-horizon effects), $|\mathcal{J}| = 3{,}360$ groups (56 arrival times, 3 age groups and 20 disease groups) with $|\mathcal{S}_j| = 15$ states and $|\mathcal{A}_j| = 6$ each, as well as $|\mathcal{L}| = 8$ resources (senior and junior doctors, nurses as well as beds for both G&A and CC separately). The resulting LP, while nontrivial in size, can be solved quickly and reliably on standard hardware with off-the-shelf LP solvers.

The fluid LP (4) has a natural interpretation. The decision variables $\sigma_{tj}(s)$ record for each DP group $j \in \mathcal{J}$ the expected number of DPs (j, i) that are in state $s \in \mathcal{S}_j$ in time period $t \in \mathcal{T}$, and the decision variables $\pi_{tj}(s, a)$ determine how often each admissible action $a \in \mathcal{A}_{jt}(s)$ is applied in expectation to the DPs accounted for by $\sigma_{tj}(s)$. The objective function of (4) maximizes the expected total rewards, and the first four constraints (in order) ensure that σ_{1j} adheres to the initial state distributions q_j, the one-step evolutions respect the transition probabilities p_{jt} as well as the selected policy π_{tj}, the resource constraints are met and that no DPs 'disappear' in any time period, respectively.

3.2 Approximation Guarantees and Feasible Policies

By construction, the fluid LP (4) constitutes a relaxation of the grouped weakly coupled DP in the sense that the optimal objective value of (4) bounds the largest achievable expected total reward in the grouped weakly coupled DP from above, and a feasible solution to (4) does not necessarily correspond to a feasible policy for the grouped weakly coupled DP. We now investigate how tight the approximation offered by the fluid LP (4) is, and how optimal solutions to (4) allow us to construct near-optimal policies for the grouped weakly coupled DP.

Recall that a deterministic policy $\pi = \{\pi_t\}_t$, $\pi_t : \mathcal{S} \to \mathcal{A}$, for the grouped weakly coupled DP assigns an action $a \in \mathcal{A}$ to each possible state $s \in \mathcal{S}$ in each time period $t \in \mathcal{T}$. We now utilize an optimal solution $(\sigma^{\text{LP}}, \pi^{\text{LP}})$ to the fluid LP (4) to construct a *randomized* policy $\pi^\star = \{\pi^\star_t\}_t$, $\pi^\star_t : \mathcal{S} \to \Delta(\mathcal{A})$, for the grouped weakly coupled DP that assigns a probability distribution over all actions $a \in \mathcal{A}$ to each possible state $s \in \mathcal{S}$ for each time period $t \in \mathcal{T}$ as follows:

$$[\pi^\star_t(s)](a) \;=\; \prod_{j \in \mathcal{J}} \prod_{i=1}^{n_j} \frac{\pi^{\text{LP}}_{tj}(s_{(j,i)}, a_{(j,i)})}{\sigma^{\text{LP}}_{tj}(s_{(j,i)})} \qquad \forall t \in \mathcal{T},\ \forall (s, a) \in \mathcal{S} \times \mathcal{A} \tag{5}$$

Here, we adopt the convention that $0/0 = 0$. Intuitively speaking, $\pi^\star$ considers each constituent DP $(j, i) \in \mathcal{J} \times \{1, \ldots, n_j\}$ independently, and it employs each action $a_{(j,i)} \in \mathcal{A}_{(j,i)}$ with a probability that equals the fraction of times this action is selected for DPs in group j that reside in the same state $s_{(j,i)}$ as the DP (j, i) at time period t in the fluid LP (4). One readily verifies from the constraints of the fluid LP (4) that for all $t \in \mathcal{T}$ and $s \in \mathcal{S}$, the policy $\pi^\star_t$ indeed defines a probability distribution over $\mathcal{A}$.

Under $\pi^\star$, the grouped weakly coupled DP evolves as follows. The initial state $\tilde{s}_1$ is random and satisfies $\mathbb{P}[\tilde{s}_1 = s] = q(s)$ for $s \in \mathcal{S}$. For $t \in \mathcal{T} \backslash \{T\}$, the transitions are governed by

$$\mathbb{P}\left[\tilde{s}_{t+1} = s'\right] \;=\; \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} [\pi^\star_t(s)](a) \cdot p_t(s' \,|\, s, a) \cdot \mathbb{P}\left[\tilde{s}_t = s\right] \qquad \forall s' \in \mathcal{S}.$$

The expected total reward of $\pi^\star$ is $\mathbb{E}\left[\sum_{t \in \mathcal{T}} \sum_{a \in \mathcal{A}} [\pi^\star_t(\tilde{s}_t)](a) \cdot r_t(\tilde{s}_t, a)\right]$.

We now study the performance of $\pi^\star$ in the grouped weakly coupled DP. Our results make use of the random state $\tilde{s}_{t,(j,i)}$ of the (j, i)-th DP and the random action $\tilde{a}_{t,(j,i)}$ applied to this DP in time period t, respectively. Recall that $N = \sum_{j \in \mathcal{J}} n_j$ denotes the total number of constituent DPs in the grouped weakly coupled DP.

Theorem 1 (Infeasible Randomized Policy; Expected Total Reward)**.** *Denote by $\theta^\star$ and θ^{DP} the expected total reward of the randomized policy $\pi^\star$ and an optimal policy for the grouped weakly coupled DP, respectively. We then have*

$$\theta^\star \;\geqslant\; \theta^{\text{DP}}$$

as well as, with probability at least $1 - |\mathcal{T}| \cdot |\mathcal{L}| \,/\, N^2$,

$$\sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \;\leqslant\; b_{tl} \;+\; \sqrt{N \log N} \cdot \max_{j \in \mathcal{J}} \, \|c_{tlj}\|_2 \qquad \forall t \in \mathcal{T},\; \forall l \in \mathcal{L},$$

where the 2-norm is taken over the components $c_{tlj}(s, a)$, $(s, a) \in \mathcal{S}_j \times \mathcal{A}_j$, of c_{tlj}.

Theorem 2 (Infeasible Randomized Policy; Worst-Case Total Reward)**.** *Denote by $\tilde{\theta}^\star$ the random total reward of the randomized policy $\pi^\star$ and by θ^{DP} the expected total reward of an optimal policy for the grouped weakly coupled DP, respectively. Then, jointly with probability at least $1 - |\mathcal{T}|(|\mathcal{L}|+1)/N^2$,*

$$\tilde{\theta}^\star \;\geqslant\; \theta^{\text{DP}} \;-\; |\mathcal{T}| \cdot \sqrt{N \log N} \cdot \max_{\substack{t \in \mathcal{T}, \\ j \in \mathcal{J}}} \, \|r_{jt}\|_2\,,$$

where the 2*-norm is taken over the components* $r_{jt}(s,a)$, $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$, *of* r_{jt}, *as well as*

$$\sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \;\leqslant\; b_{tl} \;+\; \sqrt{N \log N} \cdot \max_{j \in \mathcal{J}} \, \|c_{tlj}\|_2 \qquad \forall t \in \mathcal{T},\ \forall l \in \mathcal{L},$$

where the 2*-norm is taken over the components* $c_{tlj}(s,a)$, $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$, *of* c_{tlj}.

While Theorem 1 shows that the *expected* total reward $\theta^\star$ of the randomized policy weakly exceeds the expected total reward of an optimal policy for the grouped weakly coupled DP, Theorem 2 shows that the *actually realized* total reward $\tilde{\theta}^\star$ of the randomized policy falls significantly short of the expected total reward of an optimal policy for the grouped weakly coupled DP with low probability only. We are not aware of such worst-case bounds in the literature.

To interpret the above performance guarantees in light of our healthcare model, consider an asymptotic setting where $N \to \infty$ and where the available resources $b_{tl} \propto N$ scale in the number of patients to be treated. This corresponds to a healthcare system where the number of patients increases and where the available resources increase proportionately. In contrast, we assume that the number $|\mathcal{T}|$ of time periods, the number $|\mathcal{J}|$ of patient groups as well as the number of states $|\mathcal{S}_j|$ and actions $|\mathcal{A}_j|$ per patient group remain constant. In this setting, the probabilities $1 - |\mathcal{T}| \cdot |\mathcal{L}| \,/\, N^2$ and $1 - |\mathcal{T}|(|\mathcal{L}| + 1)/N^2$ in Theorems 1 and 2 approach 1. Since $N \cdot |\mathcal{T}| \cdot \min\{r_{jt}(s,a) \,:\, j \in \mathcal{J},\, t \in \mathcal{T} \text{ and } (s,a) \in \mathcal{S}_j \times \mathcal{A}_j\} \leqslant \theta^{\mathrm{DP}} \leqslant N \cdot |\mathcal{T}| \cdot \max\{r_{jt}(s,a) \,:\, j \in \mathcal{J},\, t \in \mathcal{T} \text{ and } (s,a) \in \mathcal{S}_j \times \mathcal{A}_j\}$, the expected total reward θ^{DP} of the grouped weakly coupled DP under any optimal policy scales with $N \cdot |\mathcal{T}|$. Since $\max_{j \in \mathcal{J}} \|c_{tlj}\|_2$ and $\max_{t \in \mathcal{T},\, j \in \mathcal{J}} \|r_{jt}\|_2$ remain constant, the resource violations grow according to $\sqrt{N \log N}$ and hence sublinearly in the number N of patients, that is, they vanish on a relative scale. The same applies for the suboptimality in the case of the worst-case bound, which however involves the additional factor $|\mathcal{T}|$. The results also reveal the price to be paid when moving from a guarantee in expectation (*cf.* Theorem 1) to a guarantee in terms of the worst case (*cf.* Theorem 2): while the expected total reward $\theta^\star$ of the randomized policy $\pi^\star$ weakly exceeds the optimal expected total reward θ^{DP}, with high probability the worst-case total reward $\tilde{\theta}^\star$ falls short of θ^{DP} by a quantity that only grows sublinearly in the number N of patients. Note that in our application, the number $|\mathcal{T}|$ of time periods impacts the number $|\mathcal{J}|$ of patient groups as patients arriving at different times are assigned to different patient groups, even if they are otherwise homogeneous. Thus our approximation guarantees depend on the selected time granularity (*e.g.*, weekly vs. daily time periods). We note that the bounds of Theorems 1 and 2, as well as the results below that build upon those theorems, can be refined by replacing $\max_{j \in \mathcal{J}} \|c_{tlj}\|_2$ and $\max_{t \in \mathcal{T}, j \in \mathcal{J}} \|r_{jt}\|_2$ with $\max_{j \in \mathcal{J}} \|c_{tlj} - \hat{c}_{tlj}\|_2$ and $\max_{t \in \mathcal{T}, j \in \mathcal{J}} \|r_{jt} - \hat{r}_{jt}\|_2$, where $\hat{c}_{tlj}$ and $\hat{r}_{jt}$ can be arbitrary constants, such as the means of the costs $\hat{c}_{tlj} = \frac{1}{|\mathcal{S}_j| \cdot |\mathcal{A}_j|} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} c_{tlj}(s, a)$ and expected rewards $\hat{r}_{jt} = \frac{1}{|\mathcal{S}_j| \cdot |\mathcal{A}_j|} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s, a)$, respectively. Indeed, Lemma 2 in the appendix—which provides the foundation of both theorems—considers this more general formulation.

The randomized policy constructed thus far will likely violate the resource constraints. As long as the resource violations are small, as is guaranteed to be the case with high probability by Theorems 1 and 2, they are likely to be acceptable in our healthcare application since the resource limits are themselves uncertain and subject to measurement inaccuracies, and the resources in our healthcare application tend to be, within certain limits, expandable. That said, we next construct a near-optimal randomized policy $\pi^{\mathrm{F}} = \{\pi_t^{\mathrm{F}}\}_t$, $\pi_t^{\mathrm{F}} : \mathcal{S} \to \Delta(\mathcal{A})$, for the grouped weakly coupled DP that is guaranteed to be feasible. To this end, we first reduce the resource limits b_{tl} in the fluid LP to

$$b_{tl}^{\mathrm{red}} = \left[b_{tl} - \sqrt{N \log N} \cdot \max_{j \in \mathcal{J}} \|c_{tlj}\|_2 \right]_+ \qquad \forall t \in \mathcal{T},\ \forall l \in \mathcal{L}, \tag{6}$$

where $[x]_+ = \max\{x,\, 0\}$. We then construct an auxiliary policy π^{red} from an optimal solution to the fluid LP with the updated resource limits b_{tl}^{red} as in equation (5). Theorems 1 and 2 guarantee that π^{red} satisfies all resource constraints across all time periods with probability at least $1 - |\mathcal{T}| \cdot |\mathcal{L}| \,/\, N^2$. In order to construct a policy π^{F} that is feasible *with certainty*, we require for each time period $t \in \mathcal{T}$ and each state $s \in \mathcal{S}$ of the grouped weakly coupled DP the existence of an action $a \in \mathcal{A}_t^{\mathrm{C}}(s)$ that satisfies the original resource constraints with right-hand sides b_{tl}. A sufficient condition is the existence of designated 'do-nothing' actions as per the following assumption.

Assumption 2 ('Do-Nothing' Actions)**.** *We assume that for each DP group $j \in \mathcal{J}$ there is a 'do-nothing' action $a_j^0 \in \mathcal{A}_j$ that is feasible (although, possibly, heavily penalized) in every state $s_j \in \mathcal{S}_j$ and that does not consume any resources.*

In the context of our healthcare application, the do-nothing action could imply that a waiting patient has to wait for another week for her treatment, or that a patient currently under treatment is sent home. We then construct the feasible policy π^{F} from the auxiliary policy π^{red} as follows: in each time period $t \in \mathcal{T}$ and for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ with positive probability $[\pi_t^{\mathrm{red}}(s)](a) >$

0, if action a violates the resource limits $\{b_{tl}\}_{t,l}$, then we shift the probability mass assigned to action a to any alternative actions a' that satisfy them. If do-nothing actions (*cf.* Assumption 2) exist, then the alternative actions can be constructed, for example, by assigning do-nothing actions to random patients until the original resource limits b_{tl} are met.

Proposition 3 (Feasible Randomized Policy; Expected Total Reward)**.** *Denote by θ^{F} and θ^{DP} the expected total reward of the feasible randomized policy π^{F} and an optimal policy for the grouped weakly coupled DP, respectively. We then have*

$$\theta^{\mathrm{F}} \;\geqslant\; \theta^{\mathrm{DP}} \;-\; \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t\in\mathcal{T},\, l\in\mathcal{L},\\ j\in\mathcal{J}}} \left\{ \frac{N}{b_{tl}} \left\| c_{tlj} \right\|_2 \right\} \cdot \left(\theta^{\mathrm{DP}} - N \cdot |\mathcal{T}| \cdot \underline{r}\right) \;-\; \frac{|\mathcal{T}|^2 \cdot |\mathcal{L}| \cdot (\overline{r} - \underline{r})}{N},$$

where the 2-norm is taken over the components $c_{tlj}(s,a)$, $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$, of c_{tlj}, $\underline{r} = \min\{r_{jt}(s,a) : j \in \mathcal{J},\, t \in \mathcal{T}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j\}$ and $\overline{r} = \max\{r_{jt}(s,a) : j \in \mathcal{J},\, t \in \mathcal{T}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j\}$.

Proposition 4 (Feasible Randomized Policy; Worst-Case Total Reward)**.** *Denote by $\tilde{\theta}^{\mathrm{F}}$ the random total reward of the feasible randomized policy π^{F} and by θ^{DP} the expected total reward of an optimal policy for the grouped weakly coupled DP, respectively. We then have*

$$\begin{aligned}\tilde{\theta}^{\mathrm{F}} \;\geqslant\; \theta^{\mathrm{DP}} \;&-\; \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t\in\mathcal{T},\, l\in\mathcal{L},\\ j\in\mathcal{J}}} \left\{ \frac{N}{b_{tl}} \left\| c_{tlj} \right\|_2 \right\} \cdot \left(\theta^{\mathrm{DP}} - N \cdot |\mathcal{T}| \cdot \underline{r}\right) \\ &-\; |\mathcal{T}| \cdot \sqrt{N \log N} \cdot \max_{t\in\mathcal{T},\, j\in\mathcal{J}} \left\| r_{jt} \right\|_2\end{aligned}$$

with probability at least $1 - |\mathcal{T}| \cdot (|\mathcal{L}| + 1)/N$, where we use the notation of Proposition 3 and where the second 2-norm is taken over the components $r_{jt}(s,a)$, $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$, of r_{jt}.

Note that the expected total reward in Proposition 3 decreases by

$$\sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t\in\mathcal{T},\, l\in\mathcal{L},\\ j\in\mathcal{J}}} \left\{ \frac{N}{b_{tl}} \left\| c_{tlj} \right\|_2 \right\} \cdot \left(\theta^{\mathrm{DP}} - N \cdot |\mathcal{T}| \cdot \underline{r}\right) \;+\; \frac{|\mathcal{T}|^2 \cdot |\mathcal{L}| \cdot (\overline{r} - \underline{r})}{N}$$

when compared to its infeasible counterpart from Theorem 1. The first term upper bounds the reduction in the optimal value of the fluid LP (4) caused by reducing the resource limits from $\{b_{tl}\}_{t,l}$ to $\{b_{tl}^{\mathrm{red}}\}_{t,l}$, whereas the second term upper bounds the reduction in expected total reward caused by moving from the infeasible auxiliary policy π^{red} to the feasible randomized policy π^{F}. In contrast,

	expected case		
	suboptimality	violation	probability
infeasible policy $\pi^\star$	0	$\sqrt{N \log N}$	$\frac{\lvert\mathcal{T}\rvert \cdot \lvert\mathcal{L}\rvert}{N^2}$
feasible policy π^{F}	$\lvert\mathcal{T}\rvert \cdot \sqrt{N \log N} + \frac{\lvert\mathcal{T}\rvert^2 \cdot \lvert\mathcal{L}\rvert}{N}$	0	0

	worst case		
	suboptimality	violation	probability
infeasible policy $\pi^\star$	$\lvert\mathcal{T}\rvert \cdot \sqrt{N \log N}$	$\sqrt{N \log N}$	$\frac{\lvert\mathcal{T}\rvert(\lvert\mathcal{L}\rvert + 1)}{N^2}$
feasible policy π^{F}	$\lvert\mathcal{T}\rvert \cdot \sqrt{N \log N}$	0	0

Table 1. Summary of the suboptimality in terms of total reward, the resource violation as well as the violation probability from Theorems 1 and 2 as well as Propositions 3 and 4 when the number N of patients grows.

the worst-case total reward in Proposition 4 only decreases by

$$\sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T},\, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \, \|c_{tlj}\|_2 \right\} \cdot \left(\theta^{\mathrm{DP}} - N \cdot |\mathcal{T}| \cdot \underline{r}\right)$$

when compared to its infeasible counterpart from Theorem 2. Indeed, while the event that the auxiliary policy π^{red} is infeasible must be accounted for in the *expected* total reward bound (*cf.* Proposition 3), it can be disregarded in the consideration of the *random* total reward (*cf.* Proposition 4) since it occurs with low probability, and the bound on the random total reward is only guaranteed to hold with high probability (as opposed to almost surely).

In order to judge the extent to which the policy π^{F} sacrifices optimality in favor of feasibility, consider again an asymptotic setting where $N \to \infty$ as well as $b_{tl} \propto N$. Since $N \cdot |\mathcal{T}| \cdot \underline{r} \leqslant \theta^{\mathrm{DP}} \leqslant N \cdot |\mathcal{T}| \cdot \overline{r}$, we have $\theta^{\mathrm{DP}} \propto N \cdot |\mathcal{T}|$. Thus, the additional decrease in expected total reward (*cf.* Proposition 3) scales according to $|\mathcal{T}| \cdot \sqrt{N \log N} + |\mathcal{T}|^2 \cdot |\mathcal{L}| \, / \, N$, whereas the additional decrease in worst-case total reward (*cf.* Proposition 4) grows according to $|\mathcal{T}| \cdot \sqrt{N \log N}$. In conclusion, the suboptimality of π^{F} continues to grow sublinearly in N, and π^{F} is asymptotically optimal as well. Table 1 summarizes the performance guarantees of the policies $\pi^\star$ and π^{F} offered by Theorems 1 and 2 as well as Propositions 3 and 4.

Remark 1. *Our construction of the feasible policy* π^{F} *relies on reducing the available resources from*

$\{b_{tl}\}_{t,l}$ *to* $\{b^{\text{red}}_{tl}\}_{t,l}$. *This construction is inspired by the work of Zayas-Cabán et al. (2019), who apply a similar idea in the context of a restless bandit problem where all bandits are governed by the same transition kernel. The main differences are that Zayas-Cabán et al. (2019) consider a single class of bandits (whereas we model multiple groups of patients), and they consider a single resource constraint for each time period that imposes an upper bound on the number of applied actions (whereas we allow for multiple resource constraints per time period that can weigh actions differently).*

Remark 2. *If the number* N *of constituent DPs in the grouped weakly coupled DP is small or moderate, then the resource reduction from* $\{b_{tl}\}_{t,l}$ *to* $\{b^{\text{red}}_{tl}\}_{t,l}$ *can be significant, and the feasible randomized policy* π^{F} *may under-utilize the factually existing resources* $\{b_{tl}\}_{t,l}$. *In such cases, a better performance may be obtained if we modify the feasible randomized policy* π^{F} *by shifting probability mass from actions that under-utilize the factual resource limits* $\{b_{tl}\}_{t,l}$ *to other actions that 'make better use' of the available resources. Here, the interpretation of a 'better use' of resources is heuristic and dependent on the application: we will outline a simple heuristic in the context of our synthetic experiment (Section 4), whereas we find that our feasible randomized policy* π^{F} *does not require any adaptation in the context of our healthcare case study (Section 5). Note that the modified policy outlined in this remark no longer enjoys the* a priori *guarantees of Propositions 3 and 4; one can nevertheless compute its performance on a specific problem instance* a posteriori *using simulation.*

4 Synthetic Experiment

To obtain further insights into our randomized policy from Section 3.2, we compare our policy with some standard weakly coupled DP policies from the literature on a stylized problem. To this end, consider a multi-armed bandit problem with a finite time horizon $T \in \{5, 25, 50\}$ where $N \in \{10{,}000, 100{,}000, 1{,}000{,}000\}$ bandits are spread equally across $J = 10$ different groups. Across all groups $j \in \mathcal{J}$, each bandit has the same number of states $|\mathcal{S}_j| \in \{2, 3, 4, 5\}$ and actions $|\mathcal{A}_j| = 2$ ('pull arm' and 'do nothing'). For each group j, we sample the initial state probabilities q_j as well as the (non-stationary) transition probabilities p_{jt} under the 'pull arm' action uniformly at random, whereas each bandit deterministically resides in its current state under the 'do nothing' action. Likewise, the rewards r_{jt} are state-dependent and drawn uniformly at random from the interval $[0, 1]$ under the 'pull arm' action, whereas they are zero under the 'do nothing' action. We have a single resource constraint which requires that in each time period at most $b \in \{1\% \cdot N,\ 5\% \cdot N,\ 10\% \cdot$

N, $25\% \cdot N\}$ arms can be pulled. The goal is to find a policy that decides which arms to pull in each time period so as to maximize the expected total reward. This problem is known in the literature as the *ordinary* (or *regular*) multi-armed bandit problem, as opposed to the *restless* multi-armed bandit problem where inactive arms passively transition to new states. While the optimal policy is known for the ordinary multi-armed bandit problem in the discounted infinite horizon case with $b = 1$ (this is the Gittins index), the problem is known to be hard in our finite horizon setting where multiple arms can be pulled in every time stage. Note that our bandit problem is characterized by a large number of bandits, each equipped with a small number of states and actions, that are spread across a relatively small number of groups and subjected to a fairly tight resource constraint. These properties are reminiscent of our healthcare case study in Section 5.

We compare our feasible randomized policy from Section 3.2 (with the adaptation outlined in Remark 2) against the Lagrangian relaxation-based feasible policy of Brown and Smith (2020) as well as the approximate LP-based feasible policy of Adelman and Mersereau (2008). Details of the three formulations are relegated to the appendix. In all three approaches, we exploit the group structure inherent in the bandit problem, which implies that the associated LP formulations scale in the number of groups (as opposed to the number of bandits) and can thus be solved within seconds with standard solvers. We report the suboptimality of all three approaches relative to the optimal value of the Lagrange relaxation (which itself overestimates the achievable expected total reward).

Table 2 compares the performance of the three policies. All results are reported as averages over 100 randomly constructed test instances, where the expected total reward of each feasible policy is computed based on 10,000 sample paths. The results show that the Lagrangian relaxation-based feasible policy of Brown and Smith (2020) and our feasible randomized policy perform quite similarly, where The approximate LP-based feasible policy of Adelman and Mersereau (2008) is not competitive on the considered instances. Further investigations revealed that this is due to the fact that an unrealistically large number of constraints would have to be sampled for the LP-based relaxation to provide reasonably accurate approximations of the true reward to-go.

5 Elective Care Scheduling in England

In this section, we apply our framework to the COVID pandemic management in England. In particular, we discuss the overall setup of our case study (Section 5.1), the employed data sources

			10,000 bandits				100,000 bandits				1,000,000 bandits			
			1%	5%	10%	25%	1%	5%	10%	25%	1%	5%	10%	25%
2 states	$T = 5$	RP	0.00%	0.10%	0.06%	0.11%	0.00%	0.02%	0.02%	0.19%	0.00%	0.00%	0.01%	0.09%
		B&S	0.01%	0.02%	0.02%	0.03%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.09%	0.51%	1.06%	3.25%	0.04%	0.55%	1.15%	3.26%	0.03%	0.38%	1.09%	3.25%
2 states	$T = 25$	RP	0.09%	0.26%	0.17%	0.22%	0.01%	0.07%	0.05%	0.12%	0.01%	0.02%	0.02%	0.10%
		B&S	0.02%	0.04%	0.03%	0.03%	0.01%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.48%	2.10%	3.92%	7.27%	0.12%	1.61%	3.01%	6.46%	0.19%	1.57%	3.02%	6.39%
2 states	$T = 50$	RP	0.10%	0.29%	0.19%	0.24%	0.03%	0.09%	0.06%	0.11%	0.01%	0.03%	0.02%	0.10%
		B&S	0.02%	0.05%	0.04%	0.03%	0.01%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	1.13%	3.65%	4.93%	8.23%	0.36%	2.39%	4.02%	8.24%	0.27%	2.26%	4.23%	8.36%
3 states	$T = 5$	RP	0.03%	0.10%	0.21%	0.16%	0.00%	0.02%	0.04%	0.05%	0.00%	0.00%	0.01%	0.02%
		B&S	0.01%	0.02%	0.03%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.04%	0.59%	1.12%	3.79%	0.02%	0.48%	1.15%	3.38%	0.03%	0.39%	1.03%	3.27%
3 states	$T = 25$	RP	0.15%	0.33%	0.29%	0.31%	0.04%	0.11%	0.09%	0.08%	0.01%	0.03%	0.03%	0.03%
		B&S	0.02%	0.04%	0.04%	0.04%	0.01%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.62%	2.58%	4.16%	7.86%	0.22%	1.72%	3.42%	7.14%	0.23%	1.62%	3.10%	7.38%
3 states	$T = 50$	RP	0.21%	0.41%	0.36%	0.26%	0.06%	0.12%	0.11%	0.09%	0.02%	0.04%	0.03%	0.04%
		B&S	0.02%	0.06%	0.05%	0.04%	0.01%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	1.13%	3.62%	5.13%	8.47%	0.40%	2.47%	4.22%	9.14%	0.33%	2.29%	4.46%	9.05%
4 states	$T = 5$	RP	0.03%	0.08%	0.19%	0.17%	0.02%	0.04%	0.03%	0.06%	0.01%	0.01%	0.01%	0.01%
		B&S	0.01%	0.02%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.09%	0.67%	1.24%	3.60%	0.10%	0.29%	0.97%	3.16%	0.08%	0.39%	0.95%	3.34%
4 states	$T = 25$	RP	0.20%	0.35%	0.35%	0.34%	0.05%	0.12%	0.10%	0.09%	0.02%	0.04%	0.04%	0.03%
		B&S	0.02%	0.04%	0.04%	0.04%	0.01%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.66%	2.41%	3.88%	7.44%	0.24%	1.64%	3.13%	6.89%	0.19%	1.44%	2.96%	7.18%
4 states	$T = 50$	RP	0.26%	0.47%	0.43%	0.33%	0.09%	0.15%	0.13%	0.09%	0.03%	0.05%	0.04%	0.03%
		B&S	0.03%	0.05%	0.05%	0.04%	0.01%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	1.30%	3.44%	5.03%	8.49%	0.41%	2.23%	4.06%	8.81%	0.35%	2.13%	4.19%	8.85%
5 states	$T = 5$	RP	0.09%	0.21%	0.15%	0.19%	0.00%	0.03%	0.05%	0.06%	0.00%	0.01%	0.01%	0.02%
		B&S	0.01%	0.02%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.10%	0.45%	0.89%	3.51%	0.07%	0.38%	0.94%	3.82%	0.04%	0.40%	1.02%	3.50%
5 states	$T = 25$	RP	0.23%	0.45%	0.44%	0.38%	0.08%	0.14%	0.13%	0.10%	0.02%	0.04%	0.04%	0.03%
		B&S	0.02%	0.04%	0.04%	0.03%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	0.69%	2.42%	3.76%	7.62%	0.31%	1.52%	3.04%	7.05%	0.22%	1.49%	2.80%	6.89%
5 states	$T = 50$	RP	0.36%	0.58%	0.48%	0.37%	0.10%	0.15%	0.16%	0.10%	0.04%	0.05%	0.05%	0.03%
		B&S	0.03%	0.05%	0.05%	0.04%	0.01%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%
		ALP	1.23%	3.53%	4.93%	8.36%	0.47%	2.19%	3.86%	8.35%	0.31%	2.08%	3.82%	8.39%

Table 2. Optimality gaps of our feasible randomized policy (RP), the Lagrangian relaxation-based feasible policy of Brown and Smith (2020) (B&S) and the feasible policy derived from the approximate LP-based relaxation (ALP), reported for different numbers of states and time horizons (rows) as well as numbers of bandits and resource restrictions (columns).

(Section 5.2), the DPs that model the patients of our healthcare model (Section 5.3) and the numerical results (Section 5.4).

5.1 Experimental Setup

As a case study, we aim to optimally schedule elective procedures for the NHS in England over a 56-week planning horizon (52 reported weeks plus an additional 4 weeks to avoid end-of-time-horizon effects), starting from March 2, 2020, with the objective of minimizing YLL. We consider a total of 10.45 million non-COVID patients that are subdivided into *(i)* electives (3.9 millions) and emergencies (6.55 millions), *(ii)* 20 disease groups and *(iii)* 3 age groups (under 25 years, 25–64 years, over 64 years). We also consider 349,279 COVID patients, all of whom are emergencies.

Our model has a weekly granularity. At the beginning of each week, a new inflow of patients in need of elective and emergency care (hereafter denoted as elective and emergency patients, respectively) enters the system. Strictly speaking, our model distinguishes between different medical procedures, and thus one and the same patient in need of several procedures is included as multiple different patients in our model. For ease of exposition, however, we continue to talk about patients in the following. Patients are then admitted to hospital, and they evolve over the duration of the week. Emergencies are always admitted to hospital upon arrival (if capacity permits). Elective patients who are not immediately admitted to hospital wait in a queue. While in the queue, an elective patient's condition might worsen and hence require emergency admission. Based on the severity of their conditions and resource availability, patients are first admitted to G&A or to CC and can transition between G&A and CC in the following weeks of hospitalization until they are eventually discharged from hospital (recovered or deceased). Transitions between G&A and CC are decided upon at the beginning of each week, and they are based on transition probabilities that are specific to the different patient groups and admission types (elective or emergency).

5.2 Data Sources

We combine a large set of modeling and administrative data to create a comprehensive dataset[4] for the NHS in England that includes elective and emergency patient inflows, transition probabilities, availability and requirement of resources. To this end, we leverage several data sources. In the

[4] The source code for the data analysis will be made available open source shortly. The data used in our case study is not freely available, but it is accessible to researchers on condition of NHS approval.

following, we indicate how these data sources map to the primitives of the fluid LP model presented in Section 3 and refer to specific sections of the accompanying paper (D'Aeth et al., 2021) for a comprehensive presentation of the data pre-processing.

We model the patient inflows, which determine the initial state probabilities q_t (for patients arriving in week 1) as well as the transition probabilities p_{jt} (for patients arriving in later weeks), via historical data (for non-COVID patients) as well as a projection method (for COVID patients). In particular, we forecast non-COVID patient inflows using individual level patient records across all hospitals in England from the Hospital Episode Statistics (HES), see NHS Digital (2020). HES contain patient level data with diagnoses, individual characteristics, care received, date of admission and time and mode of discharge from hospital. Weekly inflows of COVID patients are generated using the integrated epidemic/economic model Daedalus (Haw et al., 2020), fitting four parameters to English hospital occupancy data (NHS, 2020b) from March 20, 2020 to June 30, 2020: epidemic onset, basic reproductive number, lockdown onset, and reduction in transmission during lockdown due to pandemic mitigation strategies. The model divides the population into four age groups (0-4, 5-19, 20-64, over 65), which are then mapped into our three age groups using a linear transformation. In our numerical studies, we consider an epidemiological scenario defined by a lockdown enforced on January 1, 2021 and the maximum value of the reproductive number $R_{\max} = 1.2$ attained during the post-lockdown period. The details of our non-COVID and COVID patient inflow models are described in the Supplementary Sections 3 and 4 of D'Aeth et al. (2021), respectively.

The transition probabilities p_{jt} that characterize the evolution of hospitalized non-COVID and COVID patients are modeled with multinomial logits conditional on the length of stay in hospital using HES data and individual clinical data from patients who received care at the Imperial College Healthcare NHS Trust (Perez-Guzman et al., 2020), see Supplementary Section 5 of D'Aeth et al. (2021).

For the resources $\mathcal{L}$, staff availability is calculated from recent NHS datasets (NHS, 2020a,c), see D'Aeth et al. (2021), while staff-to-bed ratios are estimated using the Royal College of Physicians guidance (RCP London, 2020; Royal College of Nursing, 2020). As for the rewards r_{jt}, we calculate the YLL using standard life tables data provided by the Office for National Statistics (Office for National Statistics, 2019). We note that in this setting, which corresponds to the base case setting of the accompanying paper (D'Aeth et al., 2021), only the two resources 'G&A beds' and "CC beds' out of the eight considered resources are binding.

Table 3. Availability of resources and staff-to-bed ratios.

	Capacity				Staff-to-bed Ratio		
	Beds	Senior Doctors	Junior Doctors	Nurses	Senior Doctors	Junior Doctors	Nurses
G&A	102,186	10,764	8,539	43,214	15	15	5
CC	4,122	1,013	963	18,856	15	8	1

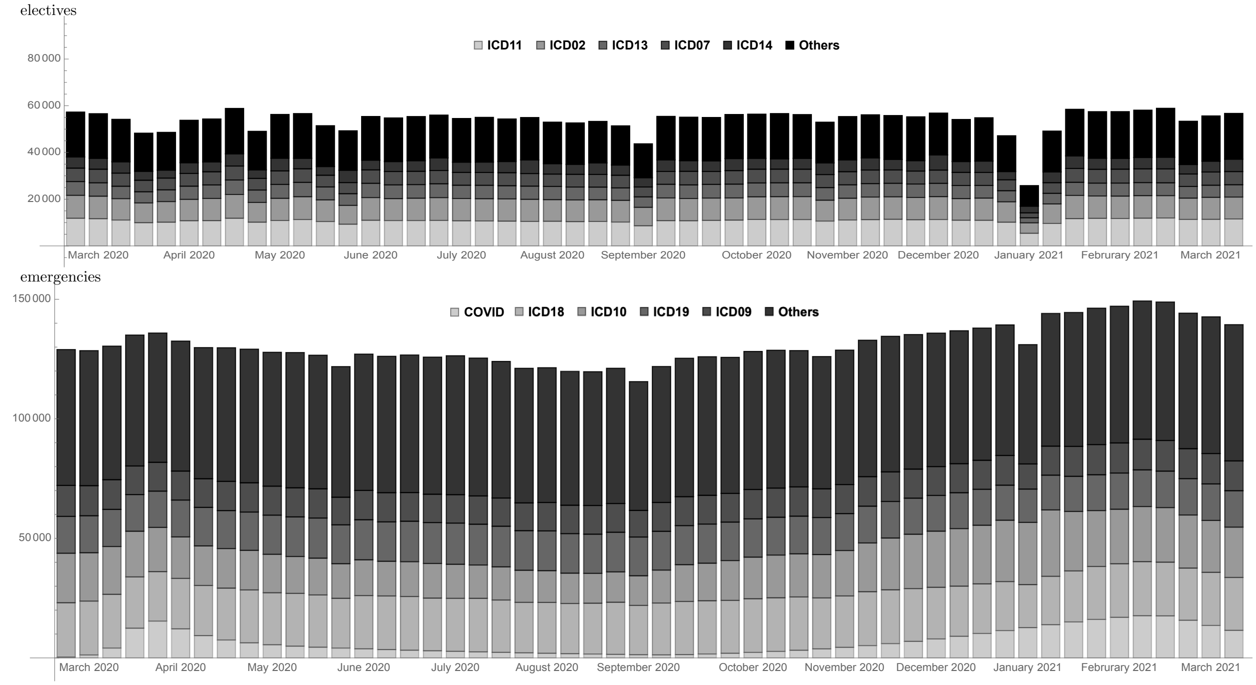

Figure 4. Weekly inflows of elective (top) and emergency (bottom) patients categorized by disease group (ICD02: neoplasms; ICD07: diseases of the eye and adnexa; ICD09: diseases of the circulatory system; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD13: diseases of the musculoskeletal system and connective tissue; ICD14: diseases of the genitourinary system; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

Table 3 and Figure 4 summarize the main input data for our case study. Table 3 reports the availability of beds and staff in G&A and CC across the NHS in England (b_{tl}), together with the corresponding staff-to-bed ratios (c_{tlj}). Figure 4 shows the weekly inflows of elective and emergency patients from March 2020 to February 2021 (q_j and p_{jt}), categorized by disease type according to the International Classification of Diseases (ICD) standard. The figure only displays the five largest patient groups individually, whereas the smaller remaining groups are collective referred to as "Others".

5.3 Dynamic Programming Model of an Individual Patient

Recall that the patients in our healthcare model are partitioned into 3,360 groups, each of which is characterized by an arrival time in the system (56 weeks), one out of 20 disease types and one out of 3 age groups. We associate a DP with each of these patient groups. All DPs share the same state and action sets, but the DPs of different patient groups differ in their admissible actions per time period and state, their initial state and transition probabilities as well their expected rewards.

Figure 5 offers a schematic representation of a patient DP. Apart from the patients that enter our healthcare model in the first week, each patient is *Dormant* until the beginning of week $t = t_0$, at which point she enters the system either as to be admitted for a planned procedure (*Elective*) or as emergency (*Emergency*). Emergency patients are admitted to hospital immediately if capacity permits; we assume that emergency patients who are denied admission die, which is represented by the *Dead* state. Elective patients that are not immediately admitted to hospital, on the other hand, remain waiting and run a risk of requiring emergency care in subsequent weeks.

Upon admission to hospital as elective or emergency, a patient requires either G&A (*Initially requires G&A* state) or CC (*Initially requires CC* state). A patient in need of G&A is admitted to G&A (*G&A* state), whereas a patient requiring CC can be assigned to either CC (*CC* state) or, in case of capacity shortages, to G&A. In the latter case, the transition into a designated G^* state implies that the patient subsequently evolves according to a different set of transition probabilities that account for an increased mortality risk associated with the denial of CC.

In the following weeks, depending on her response to the treatment, the patient can either require the same care regime or be moved to another one (*cf.* the *Requires G&A* and *Requires CC* states). Depending on resource availability, the patient then transitions between the three states *G&A*, *CC* or G^* until she eventually reaches the corresponding *Last Week* state, after which she is discharged from hospital (*Recovered* or *Dead*). We assume that a patient in a *Last Week* state only consumes half of the hospital resources, which mimics a half-week stay at the hospital. The inclusion of designated *Last Week* states allows us to account for the empirical fact that for some disease types, a large fraction of the patients require hospitalization for a few days only.

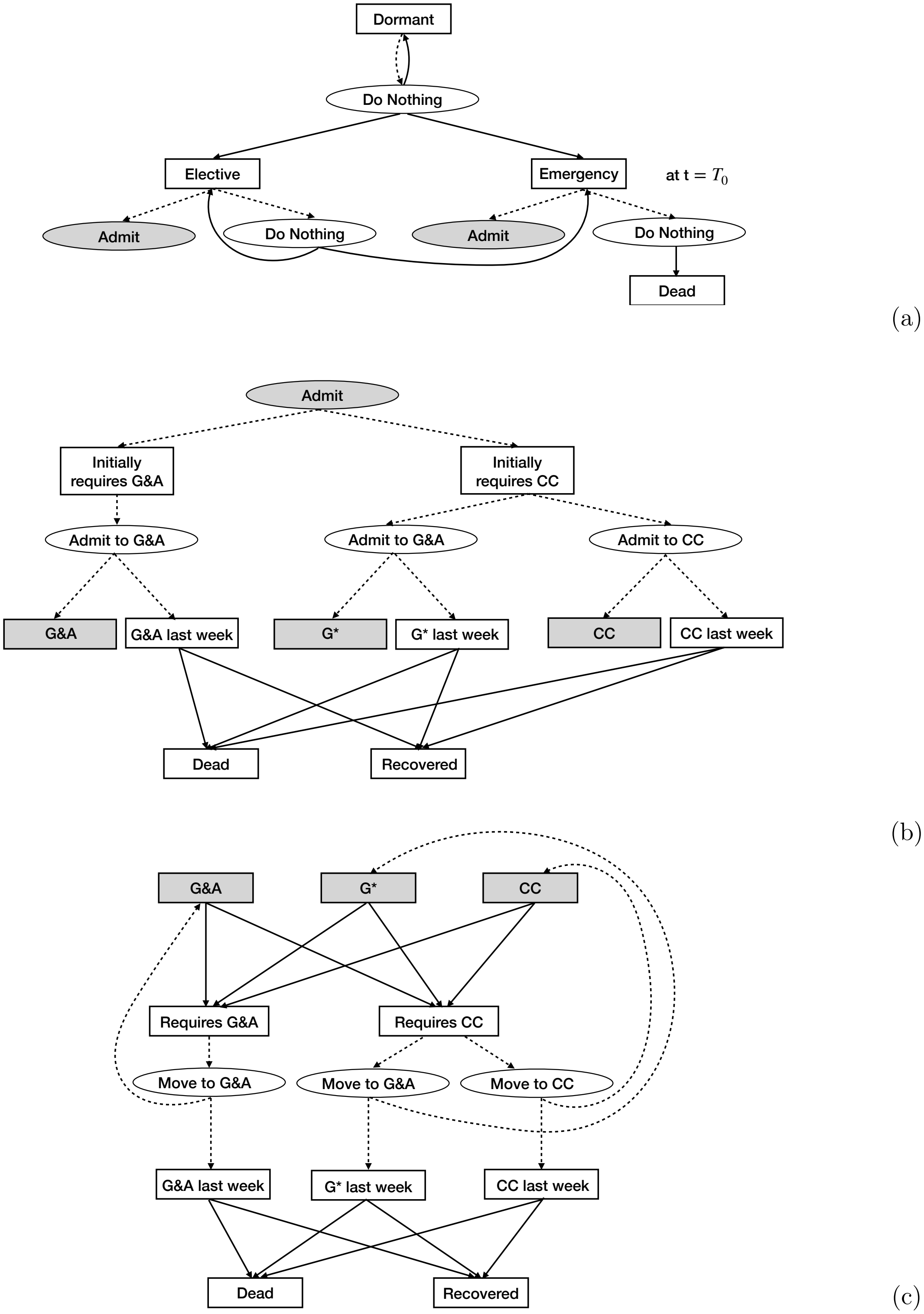

Figure 5. Schematic representation of a patient DP, from admission (a) to hospital stay (b) and discharge (c). Rectangular and oval nodes correspond to states and actions, respectively, and grey shaded nodes are exploded in the subsequent subfigure. Dotted lines correspond to immediate actions and instantaneous transitions, whereas full lines represent weekly transitions.

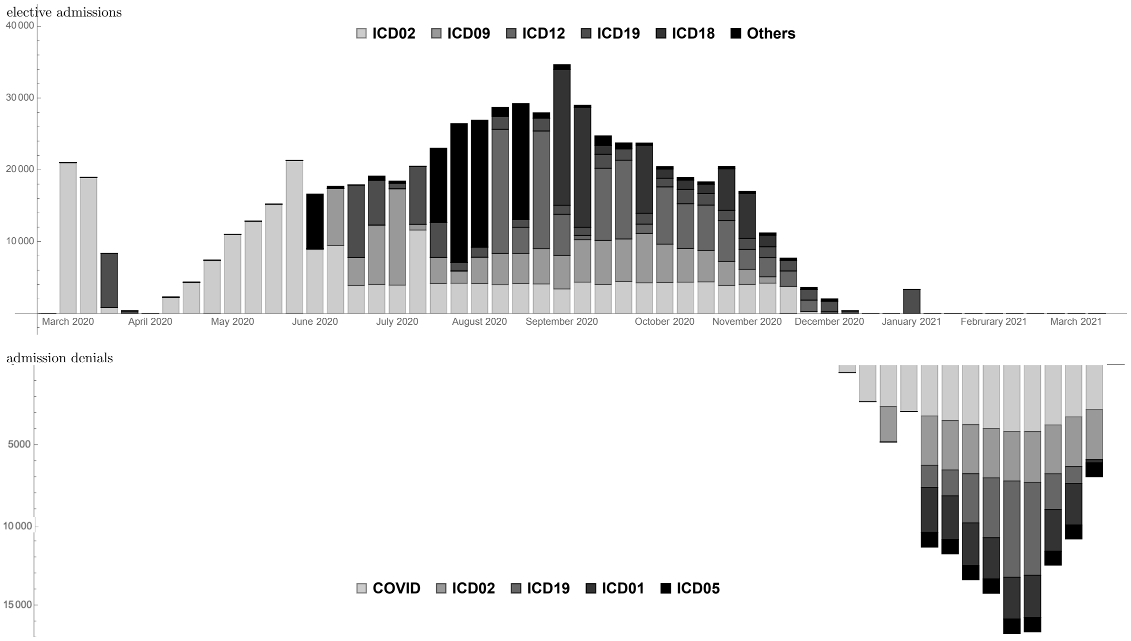

Figure 6. Weekly elective admissions (top) and admission denials (bottom) under the OS, categorized by disease group (ICD01: certain infectious and parasitic diseases; ICD02: neoplasms; ICD05: mental and behavioural disorders; ICD09: diseases of the circulatory system; ICD12: diseases of the skin and subcutaneous tissue; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

5.4 Results

We next present the numerical results of our NHS England case study (Section 5.4.1) and compare our optimized schedule (hereafter OS) against a COVID prioritization policy (hereafter CP) that resembles the one implemented in England during the pandemic (Section 5.4.2). All experiments were run on a 2.7GHz quad-core Intel i7 processor with 16GB RAM using IBM ILOG CPLEX Optimization Studio 20.1. The runtimes of the fluid LP (4) ranged between 1.5 and 2 minutes.

5.4.1 Optimized Schedule

Figure 6 shows the weekly elective admissions and emergency denials of the OS. Over the 52-week planning horizon, the OS admits to hospital 6,939,573 emergency patients (not shown), while 654,308 elective patients are admitted to hospital from week 1 to week 43 (upper part of Figure 6). Among the admitted elective patients, the most numerous groups are cancer patients (230,928), patients affected by diseases of the circulatory system (107,048) and diseases of the skin and subcutaneous tissue (101,790). In the last weeks of the planning horizon, due to the high inflow of COVID patients during the second wave of the pandemic, the available resources are insufficient to cope with the surge in demand. As a result, admission to hospital is denied to 125,346 emergency patients during weeks 38-52 (lower part of Figure 6). All of these patients are above 65 years of age, and most of them are COVID (40,962), cancer (30,069) and injury & poisoning (24,870) patients. The OS denies admission to hospital to these elderly patients as they have the lowest chances of benefiting from care.

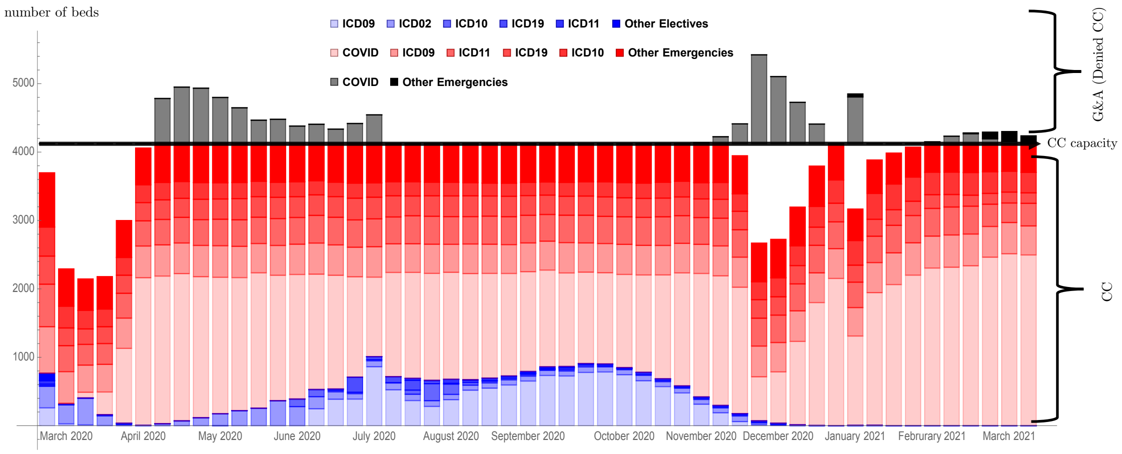

Figure 7. Weekly bed occupancy in CC by disease group (ICD02: neoplasms; ICD09: diseases of the circulatory system; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD19: injury, poisoning and certain other consequences of external causes). Patients who are denied CC, and have hence been moved to the $G^{\star}$ state, are shown above the CC capacity line.

Figure 7 shows the bed occupancy in CC. A large share of CC beds is occupied by COVID patients (40.2% average occupancy over the planning horizon), while 9.4% of the available CC beds are assigned (on average) to elective patients. During both the first and the second wave of the pandemic, capacity is insufficient, and CC is denied to some patients. The affected patients are almost exclusively COVID patients above 65 years of age, who are transferred to $G^{\star}$ due to their longer hospital stays as well as their lower capacity to benefit from treatment. G&A resources (not shown in the Figure) are fully utilized over the entire planning horizon, and the share of elective patients in G&A is on average 6.6%. Overall, the OS results in 8,233,216 YLL over the 52 weeks planning horizon, with COVID patients contributing to 64% of the YLL.

We now investigate the performance of the infeasible and feasible randomized policies from Section 3.2. Recall that the expected total reward of the infeasible randomized policy coincides with the objective value of the fluid LP and the resource violations of the infeasible randomized policy are small with high probability, whereas the expected total reward of the feasible randomized policy may fall short of the objective value of the fluid LP by a small quantity. For our case study, the infeasible randomized policy results in a YLL *de*crease of 0.01% (from 8,233,216 to 8,232,570) as well as average resource violations of 0.33% for G&A beds (ranging between 0% and 4.01% across weeks) and 1.94% for CC beds (ranging between 0% and 12.09% across weeks), respectively. The feasible randomized policy, on the other hand, is derived from the fluid LP (4) under the reduced resource limits $\{b_{tl}^{\text{red}}\}_{t,l}$ and a subsequent random assignment of 'do-nothing' actions (which amount to sending patients home; *cf.* Assumption 2) until all original resource limits $\{b_{tl}\}_{t,l}$ are met (*cf.* Section 3.2). The feasible randomized policy results in a YLL *in*crease of 1.03% (from 8,233,216 to 8,318,049), while—by construction—all resource constraints are satisfied. Since only two of the eight resources were binding, we also ran a separate experiment where we reduced the staff-to-bed ratio by 33% (reflecting the fact that the pandemic caused more stress on the system and that COVID patients required more time/attention) and assumed that up to 25% of the staff was absent during the pandemic (with the absenteeism varying weekly with the COVID numbers). Under this new experiment, junior and senior doctors in G&A as well as junior doctors in CC became additional binding resources. In this revised experiment, the optimality gap of our randomized feasible policy increased to 4.09%. Better results could potentially be obtained if we refined our randomized feasible policy along the lines of Remark 2.

5.4.2 Comparison with COVID Prioritization Policies

The results from the previous subsection suggest that denying G&A or CC admission to COVID patients might be beneficial in case of capacity shortages. This contrasts with current practice, where many countries prioritize COVID patients to the detriment of other patients. In the following, we thus compare our OS against a CP policy that always admits COVID patients and that strongly penalizes CC denial to COVID patients in the objective. Other than that, the CP policy coincides with the OS; in particular, within the aforementioned restrictions, the CP policy optimally schedules care across all patients groups.

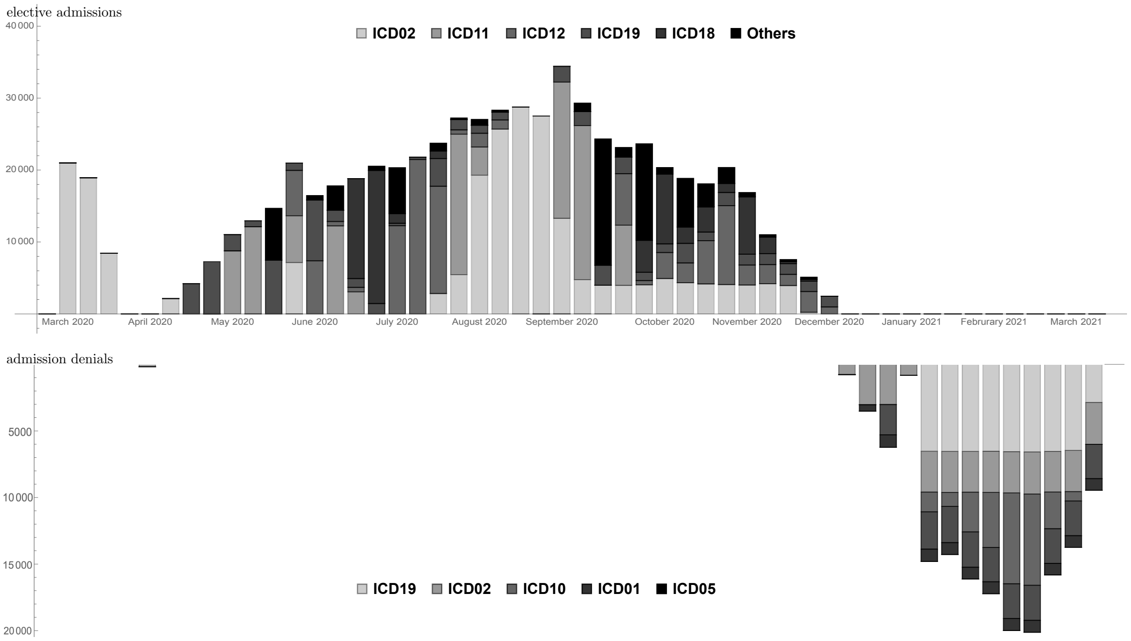

Figure 8. Weekly elective admissions (top) and admission denials (bottom) under the CP policy, categorized by disease group (ICD01: certain infectious and parasitic diseases; ICD02: neoplasms; ICD05: mental and behavioural disorders; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD12: diseases of the skin and subcutaneous tissue; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

Figure 8 shows that, while the total number of elective admissions is similar to the OS (655,415), emergency admission denials are significantly higher under the CP policy (153,092, +22.1% compared to the OS). Specifically, while all COVID patients are admitted to hospital, emergency admission is denied to patients above 65 years of age affected by injury & poisoning (55,130), cancer (35,663) and diseases of the respiratory system (26,784). The higher numbers of emergency admission denials are due to the longer treatment of COVID patients, relative to patients affected by other diseases. Under the CP policy, an average 71.6% of the CC beds are occupied by COVID patients (+75.6% compared to the OS), and the CC occupancy reaches 100% during the second wave of the pandemic (weeks 42-52). The share of electives in CC and G&A is reduced to 4.4% (-53.2% compared to the OS) and 6.5% (-1.5% compared to the OS), respectively.

Overall, the prioritization of COVID patients in admission to G&A and CC leads to an 8.7% increase in the total YLL under the CP policy compared to the OS. Figure 9 shows a breakdown of these total 719,868 YLL across the different disease groups. Significant gains in years of life are seen

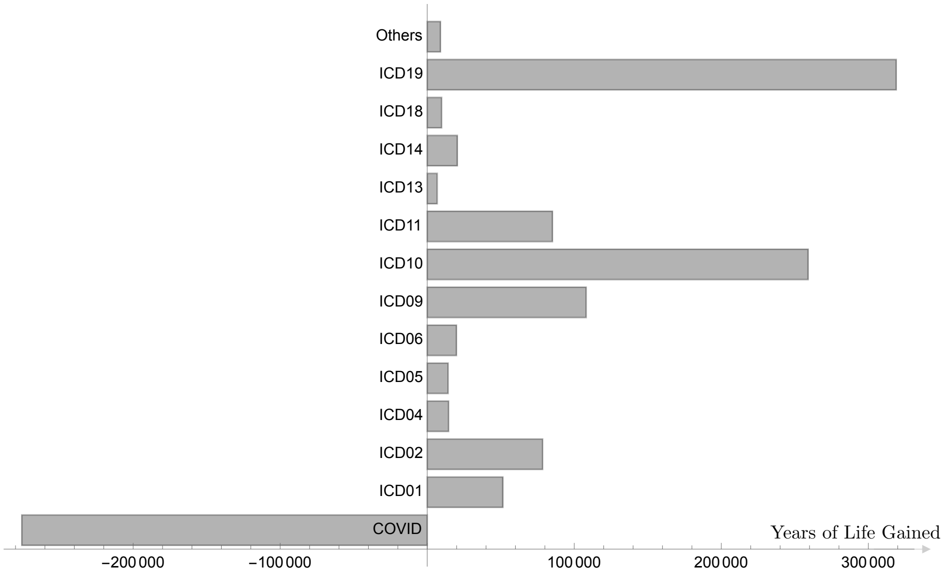

Figure 9. Years of Life Gained (*i.e.*, YLL avoided) by the OS relative to the CP policy, categorized by disease group (ICD01: certain infectious and parasitic diseases; ICD02: neoplasms; ICD04: endocrine, nutritional and metabolic diseases; ICD05: mental and behavioural disorders; ICD06: diseases of the nervous system; ICD09: diseases of the circulatory system; ICD10: diseases of the respiratory system; ICD11: diseases of the digestive system; ICD13: diseases of the musculoskeletal system and connective tissue; ICD14: diseases of the genitourinary system; ICD18: symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; ICD19: injury, poisoning and certain other consequences of external causes).

for patients affected by injury & poisoning (318,955), diseases of the respiratory system (259,012), diseases of the circulatory system (108,085), diseases of the digestive system (85,134) and cancer (78,464), to the detriment of elderly COVID patients (275,691).

6 Discussion

Our numerical case study from the previous section offers initial evidence on how the ongoing pandemic could be managed in a more effective way. That said, our model inevitably makes a number of simplifying assumptions that likely lead to a gap between the expected and the actually observed results if our optimized schedule was implemented in practice. This section reviews some of these assumptions as well as their potential impact on the real-life performance of our policy.

Disease progression and homogeneity within patient groups. With the exception of the patients in state $G^\star$ (which records a previous denial to CC), the transitions in our model—such as the weekly probability of an elective patient turning into an emergency or the weekly probability of a hospitalized patient recovering or dying—are Markovian and hence memoryless. In reality, disease progression may exhibit a more complicated dependence on waiting time, and this dependency can differ across patient groups. Non-Markovian transitions could be modeled by adding memory to the states of the patient DPs (*e.g.*, how many weeks a patient has been waiting for her surgery, and how many weeks a patient has spent in hospital), at the expense of a larger problem size as well as a more challenging parameter estimation. We also assume that patients within each group are homogenous in terms of clinical severity and disease progression. Our data do not allow us to determine whether accounting for such factors would increase or reduce the benefits of the OS compared to the CP policy.

Geographical aggregation. Our model disregards geographical differences in patient numbers, hospital resources and treatment efficacy. While this appears to be an acceptable approximation when informing the development of national guidance, a more elaborate model could subdivide the country into different regions and impose that patients can only be treated in hospitals that are sufficiently close. Computationally, a geographic disaggregation would lead to a larger and hence less tractable model. On the other hand, since the outperformance of the OS over the CP policy appears to increase when resource constraints are tight, we would expect to obtain better results in a model that imposes further constraints through a geographical disaggregation.

Data uncertainty and learning. Our model requires the estimation of a large number of input parameters, such as patient inflows, transition probabilities and resource limits, all of which are inevitably affected by significant uncertainty. The impact of this uncertainty could be analyzed further via a sensitivity analysis (as done in the accompanying paper D'Aeth et al. 2021) or by formulating and solving a robust version of the grouped weakly coupled DP. Data uncertainty impacts our model in at least two ways. Firstly, the transition probabilities for COVID patients, which are calculated based on hospital records dating between March and June 2020, likely underestimate chances of survival of COVID patients in later stages of the pandemic. The ramifications of this are not clear: on the one hand, the OS may perform better in reality since COVID patients face a higher chance of survival even outside CC, while on the other hand a better treatment in CC may imply that COVID patients should indeed be prioritized as they enjoy a higher capacity to benefit.

Secondly, our model ignores the important issue of learning: the CP policy, which aggressively prioritizes COVID patients, has led to a much better understanding of effective COVID treatments and has hence improved the transition probabilities of COVID patients, whereas our OS prioritizes non-COVID patients and hence foregoes the opportunity to learn more effective treatments. Ignoring the learning aspect likely leads to an underestimation of the benefits of the CP policy.

Health inequalities. Our model does not account for health inequalities. While YLL accounts for the principle that one year of life gained is of equal value regardless of who receives it, our model currently disregards inequity aversion in the population distribution of healthcare utilization and/or health outcomes. As an example, policy makers might be willing to sacrifice some years of life gained in favor of providing people of different age, gender, ethnicity and medical history with equal chances of survival. From a performance viewpoint, we expect our OS to fare well in this setting since equity considerations constitute additional constraints. From a tractability perspective, some of these restrictions may require the imposition of logical constraints in the fluid LP and thus result in less tractable mixed-integer linear programs. Moreover, it is unclear whether our performance guarantees for the randomized policies can be extended to this case.

Capacity and hospital acquired infections. Our model assumes that the timing and the magnitude of the patient inflows as well as the availability of staff is independent of the hospital occupancy rates. Since COVID is highly infectious, however, both non-COVID patients and hospital staff may be exposed to the virus and—in absence of protective behavior—may spread it to the community. It would therefore be instructive to study the impact of hospital occupancy rates, which are immediate consequences of the admissions decisions in our model, on hospital acquired infections, changes in care seeking behavior, ward closures as well as workload-dependent staff absenteeism (Green et al., 2013). Our model also assumes that capacity can be re-assigned between COVID and non-COVID cases on short notice. Since COVID patients require isolation and dedicated staff to reduce the risk of infections, this is not the case in practice, and our model may therefore overestimate the available capacity. We believe that this issue is attenuated by the fact that we are modeling the health system of an entire nation, as opposed to an individual hospital.

Despite the aforementioned limitations, which are to some extent inevitable when modeling complex real-world systems, our findings provide encouraging evidence that our model can help to inform strategic public health policy decisions. Specifically, our optimization framework, by considering the entire patient population as well as prioritizing care based on capacity to benefit, could contribute to an improvement of the current practice which seldom looks at the health system from a global perspective and therefore likely results in suboptimal decisions.

Acknowledgements. The authors would like to sincerely thank the department editor, the associate editor as well as the two referees for their constructive and insightful feedback that has led to substantial improvements of the manuscript.

References

Adelman, D. and A. J. Mersereau (2008). Relaxations of weakly coupled stochastic dynamic programs. *Operations Research 56*(3), 712–727.

Argenziano, M., K. Fischkoff, and C. R. Smith (2020). Surgery scheduling in a crisis. *New England Journal of Medicine 382*(23), e87.

Armony, M. and C. Maglaras (2004). On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research 52*(2), 271–292.

Bekker, R. and P. M. Koeleman (2011). Scheduling admissions and reducing variability in bed demand. *Health Care Management Science 14*(3), 237–249.

Benaïm, M. and J.-Y. Le Boudec (2008). A class of mean field interaction models for computer and communication systems. *Performance Evaluation 65*(11–12), 823–838.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*, Volume 1. Athena Scientific.

Bertsimas, D., D. Gamarnik, and J. Sethuraman (2003). From fluid relaxations to practical algorithms for high-multiplicity job-shop scheduling: The holding cost objective. *Operations Research 51*(5), 798–813.

Bertsimas, D., G. Lukin, L. Mingardi, et al. (2020). COVID-19 mortality risk assessment: An international multi-center study. *PLOS ONE 15*(12), 1–13.

Bertsimas, D. and V. V. Mišić (2016). Decomposable Markov decision processes: A fluid optimization approach. *Operations Research 64*(6), 1537–1555.

Bertsimas, D., J. Pauphilet, J. Stevens, et al. (2020). Predicting inpatient flow at a major hospital using interpretable analytics. *medRxiv 2020.05.12.20098848v2*.

Brown, D. B. and J. E. Smith (2020). Index policies and performance bounds for dynamic selection problems. *Management Science 66*(7), 3029–3050.

Brown, D. B. and J. Zhang (2020). Dynamic programs with shared resources and signals: Dynamic fluid policies and asymptotic optimality. *Available at SSRN 3728111*.

Burki, T. K. (2020). Cancer guidelines during the COVID-19 pandemic. *The Lancet Oncology 21*(5), 629–630.

Caro, F. and J. Gallien (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science 53*(2), 276–292.

Chan, C. W., V. F. Farias, N. Bambos, et al. (2012). Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research 60*(6), 1323–1341.

Christen, P., J. D'Aeth, A. Lochen, et al. (2021). The J-IDEA pandemic planner: A framework for implementing hospital provision interventions during the COVID-19 pandemic. *Medical Care, Published Ahead-of-Print*.

Ciocchetta, F., A. Degasperi, J. Hillston, and M. Calder (2009). Some investigations concerning the CTMC and the ODE model derived from Bio-PEPA. *Electronic Notes in Theoretical Computer Science 229*(1), 145–163.

D'Aeth, J., S. Ghosal, F. Grimm, et al. (2021). Optimal national prioritization policies for hospital care during the SARS-CoV-2 pandemic. *Nature Computational Science 1*, 521–531.

Davis, C., M. Gao, M. Nichols, et al. (2020). Predicting hospital utilization and inpatient mortality of patients tested for COVID-19. *medRxiv 2020.12.04.20244137*.

De Farias, D. P. and B. Van Roy (2004). On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research 29*(3), 462–478.

Déry, J., A. Ruiz, F. Routhier, et al. (2020). A systematic review of patient prioritization tools in non-emergency healthcare services. *Systematic Reviews 9*(227), 1–14.

DIVI (2020). Entscheidungen über die Zuteilung von Ressourcen in der Notfall und der Intensivmedizin im Kontext der COVID-19-Pandemie. `https://www.divi.de/joomlatools-files/docman-files/publikationen/covid-19-dokumente/200325-covid-19-ethik-empfehlung-v1.pdf`. Accessed on 1 December 2021.

Eichberg, D. G., A. H. Shah, E. M. Luther, et al. (2020). Letter: Academic neurosurgery department response to COVID-19 pandemic: the University of Miami/Jackson Memorial Hospital model. *Neurosurgery 87*(1), E63–E65.

Fujita, K., T. Ito, Z. Saito, et al. (2020). Impact of COVID-19 pandemic on lung cancer treatment scheduling. *Thoracic Cancer 11*(10), 2983–2986.

Gao, Y., G.-Y. Cai, W. Fang, et al. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature Communications 11*(1), 1–10.

Gardner, T., C. Fraser, and S. Peytrignet (2020). Elective care in England: Assessing the impact of COVID-19 and where next. Technical report, The Health Foundation.

Green, L. V., S. Savin, and N. Savva (2013). “Nursevendor problem”: Personnel staffing in the presence of endogenous absenteeism. *Management Science 59*(10), 2237–2256.

Haw, D., P. Christen, G. Forchini, et al. (2020). DAEDALUS: An economic-epidemiological model to optimize economic activity while containing the SARS-CoV-2 pandemic. Technical report, Imperial College London.

Hawkins, J. T. (2003). *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. Ph. D. thesis, Massachusetts Institute of Technology.

Helm, J. E., S. AhmadBeygi, and M. P. Van Oyen (2011). Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management 20*(3), 359–374.

Hu, W. and P. I. Frazier (2017). An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv preprint arXiv:1707.00205*.

Joebges, S. and N. Biller-Andorno (2020). Ethics guidelines on COVID-19 triage—an emerging international consensus. *Critical Care 24*(1), 201.

Kim, S.-H., C. W. Chan, M. Olivares, et al. (2015). ICU Admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science 61*(1), 19–38.

MacCormick, A. D., W. G. Collecutt, and B. R. Parry (2003). Prioritizing patients for elective surgery: A systematic review. *ANZ Journal of Surgery 73*(8), 633–642.

Maglaras, C. (2000). Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *The Annals of Applied Probability 10*(3), 897–929.

Maglaras, C. and J. Meissner (2006). Dynamic pricing strategies for multiproduct revenue management problems. *Manufacturing & Service Operations Management 8*(2), 136–148.

Marklund, J. and K. Rosling (2012). Lower bounds and heuristics for supply chain stock allocation. *Operations Research 60*(1), 92–105.

McCabe, R., N. Schmit, P. Christen, et al. (2020). Adapting hospital capacity to meet changing demands during the COVID-19 pandemic. *BMC Medicine 18*(329), 1–12.

Meng, F., J. Qi, M. Zhang, et al. (2015). A robust optimization model for managing elective admission in a public hospital. *Operations Research 63*(6), 1452–1467.

Miao, S., S. Jasin, and X. Chao (2021). Asymptotically optimal Lagrangian policies for multi-warehouse multi-store systems with lost sales. *Operations Research, Forthcoming*.

Moris, D. and E. Felekouras (2020). Surgery scheduling in a crisis: Effect on cancer patients. *Journal of BUON 25*(4), 2123–2124.

Nambiar, M., D. Simchi-Levi, and H. Wang (2021). Dynamic inventory allocation with demand learning for seasonal goods. *Production and Operations Management 30*(3), 750–765.

Negopdiev, D., C. Collaborative, and E. Hoste (2020). Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. *British Journal of Surgery 107*(11), 1440–1449.

NHS (2020a). Bed Availability and Occupancy. `https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/`. Accessed on 1 December 2021.

NHS (2020b). COVID-19 Hospital Activity. `https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-hospital-activity/`. Accessed on 1 December 2021.

NHS (2020c). Critical Care Bed Capacity and Urgent Operations Cancelled. `https://www.england.nhs.uk/statistics/statistical-work-areas/critical-care-capacity/`. Accessed on 1 December 2021.

NHS (2020d). Important and Urgent: Next Steps on NHS Response to COVID-19. `https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/03/urgent-next-steps-on-nhs-response`
`-to-covid-19-letter-simon-stevens.pdf`. Accessed on 1 December 2021.

NHS Digital (2020). Hospital Episode Statistics (HES). `https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics`. Accessed on 1 December 2021.

NICE (2020). COVID-19 Rapid Guideline: Critical Care in Adults. `https://www.nice.org.uk/guidance/ng159`. Accessed on 1 December 2021.

Office for National Statistics (2019). Past and Projected Period and Cohort Life Tables, 2018-based, UK. `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/pastandprojecteddatafromtheperiodandcohortlifetables/1981to2068`. Accessed on 1 December 2021.

Ouyang, H., N. T. Argon, and S. Ziya (2020). Allocation of intensive care unit beds in periods of high demand. *Operations Research 68*(2), 591–608.

Perez-Guzman, P. N., A. Daunt, S. Mukherjee, et al. (2020). Clinical characteristics and predictors of outcomes of hospitalized patients with COVID-19 in a multi-ethnic London NHS Trust: A retrospective cohort study. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America ciaa1091*.

Phua, J., L. Weng, L. Ling, et al. (2020). Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. *The Lancet Respiratory Medicine 8*(5), 506–517.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

RCP London (2020). Safe medical staffing. `https://www.rcplondon.ac.uk/projects/outputs/safe-medical-staffing`. Accessed on 1 December 2021.

Riccioni, L., G. Bertolini, A. Giannini, et al. (2020). Raccomandazioni di etica clinica per l'ammissione a trattamenti intensivi e per la loro sospensione, in condizioni eccezionali di squilibrio tra necessità e risorse disponibili. *Recenti Progressi in Medicina 111*(4), 207–211.

Rizmie, D., M. Miraldo, R. Atun, et al. (2019). The effect of extreme temperature on emergency admissions across vulnerable populations in England: An observational study. *The Lancet 394*(Special Issue 2), S7.

Royal College of Nursing (2020). Staffing levels | Advice guides | Royal College of Nursing. `https://www.rcn.org.uk/get-help/rcn-advice/staffing-levels`. Accessed on 1 December 2021.

Shi, P., J. Helm, J. Deglise-Hawkinson, et al. (2019). Timing it right: Balancing inpatient congestion versus readmission risk at discharge. *Available at SSRN 3202975*.

Soltany, A., M. Hamouda, A. Ghzawi, et al. (2020). A scoping review of the impact of COVID-19 pandemic on surgical practice. *Annals of Medicine and Surgery 57*, 24–36.

Sud, A., M. E. Jones, J. Broggio, et al. (2020). Collateral damage: the impact on outcomes from cancer surgery of the COVID-19 pandemic. *Annals of Oncology 31*(8), 1065–1074.

The Nuffield Council on Bioethics (2020). Statement: The Need for National Guidance on Resource Allocation Decisions in the COVID-19 Pandemic. `https://www.nuffieldbioethics.org/news/statement-the-need-for-national-guidance-on-resource-allocation-decisions-in-the-covid-19-pandemic`. Accessed on 1 December 2021.

Tsitsiklis, J. N. and K. Xu (2012). On the power of (even a little) resource pooling. *Stochastic Systems 2*(1), 1–66.

Tzeng, C.-W. D., M. Teshome, M. H. Katz, et al. (2020). Cancer surgery scheduling during and after the COVID-19 first wave: the MD Anderson cancer center experience. *Annals of Surgery 272*(2), e106–e111.

Vaid, A., S. Somani, A. Russak, et al. (2020). Machine learning to predict mortality and critical events in COVID-19 positive New York City patients: A cohort study. *Journal of Medical Internet Research 22*(11), 1–19.

Yoon, D. H., S. Koller, P. M. N. Duldulao, et al. (2020). COVID-19 impact on colorectal daily practice—how long will it take to catch up? *Journal of Gastrointestinal Surgery 25*, 260–268.

Zayas-Cabán, G., S. Jasin, and G. Wang (2019). An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Advances in Applied Probability 51*(3), 745–772.

Appendix A: Implementation Details for the Synthetic Experiment

Our feasible randomized policy is based on the solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ to the fluid LP (4) with the reduced resource limits $\{b_{tl}^{\mathrm{red}}\}_{t,l}$ defined in (6). The solution $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ allows us to construct the infeasible auxiliary policy π^{red} according to equation (5). We next construct the feasible randomized policy π^{F} from π^{red} as discussed in Section 3.2: In each time period $t \in \mathcal{T}$ and for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ with positive probability $[\pi_t^{\mathrm{red}}(s)](a) > 0$, if action a violates the budget b then we shift the probability mass to an action a' that 'un-pulls' arms of each bandit group $j \in \mathcal{J}$ in state $s \in \mathcal{S}_j$ in order of *ascending* probabilities $\pi_{tj}^{\mathrm{LP}}(s, \text{'pull arm'})/\sigma_{tj}^{\mathrm{LP}}(s)$ as indicated by equation (5). Likewise, as pointed out in Remark 2, if action a under-utilizes the budget b then we shift the probability mass to an action a' that pulls additional arms of each bandit group $j \in \mathcal{J}$ in state $s \in \mathcal{S}_j$ in order of *descending* probabilities $\pi_{tj}^{\mathrm{LP}}(s, \text{'pull arm'})/\sigma_{tj}^{\mathrm{LP}}(s)$ as indicated by equation (5). In other words, bandits of group $j \in \mathcal{J}$ in state $s \in \mathcal{S}_j$ are more (less) likely to be activated in the final policy than in π^{F} if their associated probabilities $\pi_{tj}^{\mathrm{LP}}(s, \text{'pull arm'})/\sigma_{tj}^{\mathrm{LP}}(s)$ are higher (lower). Any possible ties are broken according to the bandit group index $j \in \mathcal{J}$ and the index $i \in \{1, \ldots, n_j\}$ of the bandit within the group.

The Lagrangian relaxation-based feasible policy of Brown and Smith (2020) utilizes the Lagrangian relaxation (3) dual to our fluid LP (4). For each time period $t \in \mathcal{T}$, bandit group $j \in \mathcal{J}$ and bandit state $s \in \mathcal{S}_j$, we approximate the expected excess reward of pulling an arm via

$$f_{tj}(s) \;=\; r_{jt}(s, \text{'pull arm'}) + \sum_{s' \in \mathcal{S}_j} p_{jt}(s' \mid s, \text{'pull arm'}) \cdot V_{t+1,j}^{\lambda}(s') - V_{t+1,j}^{\lambda}(s) + \epsilon \cdot \mathcal{B}\left[\frac{\pi_{tj}^{\mathrm{LP}}(s, \text{'pull arm'})}{\sigma_{tj}^{\mathrm{LP}}(s)}\right],$$

where V^{λ} and $(\sigma^{\mathrm{LP}}, \pi^{\mathrm{LP}})$ denote optimal solutions to the primal Lagrangian relaxation (3) and the dual fluid LP (4), respectively, ϵ is a sufficiently small positive constant (we choose $\epsilon = 10^{-5}$ in our numerical experiments), and $\mathcal{B}$ denotes a Bernoulli random variable that takes value 1 with the probability given as its argument. The first three terms in $f_{tj}(s)$ approximate the difference in expected total reward between activating a bandit of group $j \in \mathcal{J}$ that is in state $s \in \mathcal{S}_j$ at time $t \in \mathcal{T}$ and not activating the bandit, whereas the last term is a tiebreaker that leverages the probability of pulling the arm under the fluid LP solution. In each time period, the Lagrangian relaxation-based feasible policy of Brown and Smith (2020) only considers arms whose corresponding expected excess rewards $f_{tj}(s)$ are non-negative, and it pulls up to b of those arms in order of descending expected excess rewards (less than b arms are pulled if all other arms have negative expected excess rewards). Any possible ties are broken according to the bandit group index $j \in \mathcal{J}$ and the index $i \in \{1, \ldots, n_j\}$ of the bandit within the group.

The approximate LP-based feasible policy of Adelman and Mersereau (2008), finally, considers the following LP approximation of the grouped weakly coupled DP,

$$\begin{aligned}
\underset{V}{\text{minimize}} \quad & \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} n_j \cdot q_j(s) \cdot V_{1j}(s) \\
\text{subject to} \quad & \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} V_{tj}(s_{ji}) \geqslant \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} r_{jt}(s_{ji}, a_{ji}) + \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} \sum_{s' \in \mathcal{S}_j} p_{jt}(s' \,|\, s_{ji}, a_{ji}) \cdot V_{t+1,j}(s') \\
& \qquad \forall t \in \mathcal{T},\ \forall (s, a) \in \bigtimes_{j \in \mathcal{J}} [\mathcal{S}_j^{n_j}] \times \bigtimes_{j \in \mathcal{J}} [\mathcal{A}_j^{n_j}] \;:\; (s, a) \in \mathcal{S} \times \mathcal{A}_t^{\mathrm{C}}(s),
\end{aligned}$$

where $(s, a) \in \mathcal{S} \times \mathcal{A}_t^{\mathrm{C}}(s)$ denotes the resource-feasible state-action pairs in period t. Since the number of constraints in this problem scales exponentially in the problem dimensions, we follow the constraint sampling approach of De Farias and Van Roy (2004) and randomly sample 10,000 constraints. For each time period $t \in \mathcal{T}$, bandit group $j \in \mathcal{J}$ and bandit state $s \in \mathcal{S}_j$, we approximate the expected excess reward of pulling an arm via

$$f_{tj}(s) \;=\; r_{jt}(s, \text{‘pull arm’}) \;+\; \sum_{s' \in \mathcal{S}_j} p_{jt}(s' \,|\, s, \text{‘pull arm’}) \cdot V_{t+1,j}(s') \;-\; V_{t+1,j}(s),$$

where V denotes an optimal solution the above LP approximation. In each time period, the approximate LP-based feasible policy of Adelman and Mersereau (2008) only considers arms whose corresponding expected excess rewards $f_{tj}(s)$ are non-negative, and it pulls up to b of those arms in order of descending expected excess rewards (less than b arms are pulled if all other arms have negative expected excess rewards). Any possible ties are broken according to the bandit group index $j \in \mathcal{J}$ and the index $i \in \{1, \ldots, n_j\}$ of the bandit within the group.

Appendix B: Proofs

Proof of Proposition 1. Proofs of this statement can be found, among others, in the papers of Hawkins (2003), Adelman and Mersereau (2008) and Brown and Zhang (2020). □

Proof of Proposition 2. We introduce the non-negative dual variables $\pi = \{\pi_{tj}\}_{t,j}$ with $\pi_{tj} : \mathcal{S}_j \times \mathcal{A}_j \to \mathbb{R}_+$, $t \in \mathcal{T}$ and $j \in \mathcal{J}$, and we set

$$\pi_{tj}(s,a) = 0 \qquad \forall t \in \mathcal{T},\ \forall j \in \mathcal{J},\ \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j \backslash \mathcal{A}_{jt}(s)$$

since the first constraint in the Lagrangian relaxation (3) only needs to be satisfied for the admissible actions $a \in \mathcal{A}_{jt}(s)$. The Lagrangian function is then given by

$$\begin{aligned} L(V^\lambda, \lambda; \pi) &= \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} \lambda_{tl} b_{tl} + \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} n_j \cdot q_j(s) \cdot V_{1j}^\lambda(s) + \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a) \\ &\quad \left(r_{jt}(s,a) - \sum_{l \in \mathcal{L}} \lambda_{tl} c_{tlj}(s,a) + \sum_{s' \in \mathcal{S}_j} p_{jt}(s' \mid s,a) \cdot V_{t+1,j}^\lambda(s') - V_{tj}^\lambda(s) \right) \\ &= \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a) \cdot r_{tj}(s,a) + \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} \lambda_{tl} \left(b_{tl} - \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} c_{tlj}(s,a) \cdot \pi_{tj}(s,a) \right) \\ &\quad + \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} V_{1j}^\lambda(s) \left(n_j \cdot q_j(s) - \sum_{a \in \mathcal{A}_j} \pi_{1j}(s,a) \right) \\ &\quad + \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a) \sum_{s' \in \mathcal{S}_j} p_{jt}(s' \mid s,a) \cdot V_{t+1,j}^\lambda(s') - \sum_{t \in \mathcal{T} \backslash \{1\}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a) \cdot V_{tj}^\lambda(s) \end{aligned}$$

Replacing t by $t+1$ and s by s' in the last term, we obtain

$$\sum_{t \in \mathcal{T} \backslash \{1\}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a) \cdot V_{tj}^\lambda(s) \quad = \quad \sum_{t \in \mathcal{T} \backslash \{T\}} \sum_{j \in \mathcal{J}} \sum_{s' \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{t+1,j}(s',a) \cdot V_{t+1,j}^\lambda(s').$$

Since $V^{\lambda}_{T+1,j}(s) = 0$ for all $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$, the Lagrangian function $L(V^{\lambda}, \lambda; \pi)$ simplifies to

$$\begin{aligned} L(V^{\lambda}, \lambda; \pi) \;\; = \;\; & \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a) \cdot r_{jt}(s,a) + \sum_{t=1}^{T} \sum_{l \in \mathcal{L}} \lambda_{tl} \left(b_{tl} - \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} c_{tlj}(s,a) \cdot \pi_{tj}(s,a) \right) \\ & + \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} V^{\lambda}_{1j}(s) \left(n_j \cdot q_j(s) - \sum_{a \in \mathcal{A}_j} \pi_{1j}(s,a) \right) \\ & + \sum_{t \in \mathcal{T} \setminus \{T\}} \sum_{j \in \mathcal{J}} \sum_{s' \in \mathcal{S}_j} V^{\lambda}_{t+1,j}(s') \left(\sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a) \cdot p_{jt}(s' \mid s,a) - \sum_{a \in \mathcal{A}_j} \pi_{t+1,j}(s',a) \right) \end{aligned}$$

Minimizing the Lagrangian function $L(V^{\lambda}, \lambda; \pi)$ over $V^{\lambda} = \{V^{\lambda}_{tj}\}_{t,j}$, $t \in \mathcal{T}$ and $j \in \mathcal{J}$, with $V^{\lambda}_{tj} : \mathcal{S}_j \to \mathbb{R}$ and $\lambda = \{\lambda_{tl}\}_{t,l}$, $t \in \mathcal{T}$ and $l \in \mathcal{L}$, with $\lambda_{tl} \geqslant 0$ results in the dual fluid LP (4), where we have introduced the auxiliary variables $\sigma_{tj}(s) = \sum_{a \in \mathcal{A}_j} \pi_{tj}(s,a)$ to aid exposition. Strong duality between the LPs (3) and (4) holds since both problems are feasible by construction: In (3), for example, we can choose $\lambda_{tl} = 0$ for all $t \in \mathcal{T}$ and $l \in \mathcal{L}$ as well as $V^{\lambda}_{tj}(s) = (T - t + 1) \cdot \max\{r_{jt}(s,a) : j \in \mathcal{J},\, t \in \mathcal{T},\, (s,a) \in \mathcal{S} \times \mathcal{A}\} < +\infty$, $t \in \mathcal{T}$, $j \in \mathcal{J}$ and $s \in \mathcal{S}_j$. □

The proofs of Theorems 1 and 2 rely on the following two auxiliary results, which we prove first.

Lemma 1. *The following equations hold for all $t \in \mathcal{T}$, $(j,i) \in \mathcal{J} \times \{1, \ldots, n_j\}$ and $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$:*

$$\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right] \;=\; \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{n_j}.$$

Proof of Lemma 1. According to our definition of the randomized policy $\pi^{\star}$, we have

$$\mathbb{P}\left[\tilde{a}_{t,(j,i)} = a \,|\, \tilde{s}_{t,(j,i)} = s\right] \;=\; \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{\sigma^{\mathrm{LP}}_{tj}(s)} \qquad \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j$$

for each $t \in \mathcal{T}$ and $(j,i) \in \mathcal{J} \times \{1, \ldots, n_j\}$, which in turn implies that

$$\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right] \;=\; \mathbb{P}\left[\tilde{a}_{t,(j,i)} = a \,|\, \tilde{s}_{t,(j,i)} = s\right] \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{\sigma^{\mathrm{LP}}_{tj}(s)} \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right]. \tag{7}$$

In the remainder of the proof, we show via induction on $t \in \mathcal{T}$ that $\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = \sigma^{\mathrm{LP}}_{tj}(s) \,/\, n_j$ for all $t \in \mathcal{T}$, $(j,i) \in \mathcal{J} \times \{1, \ldots, n_j\}$ and $s \in \mathcal{S}_j$, which concludes the proof.

For $t = 1$, the definition of the weakly coupled DP implies that $\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = q_j(s)$, and the first constraint of the fluid LP (4) ensures that $\sigma^{\mathrm{LP}}_{1j}(s) = n_j \cdot q_j(s)$. Assume now that $\mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] = \sigma^{\mathrm{LP}}_{tj}(s) \,/\, n_j$ for some $t \in \mathcal{T}\backslash\{T\}$. We then have

$$
\begin{aligned}
\mathbb{P}\left[\tilde{s}_{t+1,(j,i)} = s'\right] \;&=\; \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s'|s,a) \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right] \\
&=\; \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s'|s,a) \cdot \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{\sigma^{\mathrm{LP}}_{tj}(s)} \cdot \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s\right] \\
&=\; \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} p_{jt}(s'|s,a) \cdot \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{n_j} \;=\; \frac{\sigma^{\mathrm{LP}}_{t+1,j}(s')}{n_j},
\end{aligned}
$$

where the first identity follows from the definition of the weakly coupled DP, the second identity is due to (7), the third identity follows from the induction hypothesis, and the last identity is due to the second constraint of the fluid LP (4). □

Lemma 2. *Let* $\tilde{\alpha}^{\star}_{tj}(s,a) = \sum_{i=1}^{n_j} \mathbf{1}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right]$ *record the number of DPs in the* j*-th group that are in state* s *and to which action* a *is applied at time* t*. Then*

$$
\mathbb{E}\left[\tilde{\alpha}^{\star}_{tj}(s,a)\right] \;=\; \pi^{\mathrm{LP}}_{tj}(s,a) \qquad \forall t \in \mathcal{T},\ \forall j \in \mathcal{J},\ \forall (s,a) \in \mathcal{S}_j \times \mathcal{A}_j.
$$

Furthermore, with probability at least $1 - |\mathcal{T}| \,/\, N^2$*, we have*

$$
\sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} \omega_{tj}(s,a) \cdot \left[\tilde{\alpha}^{\star}_{tj}(s,a) - \pi^{\mathrm{LP}}_{tj}(s,a)\right] \;\leqslant\; \sqrt{N \log N} \cdot \max_{j \in \mathcal{J}} \left\|\omega_{tj} - \hat{\omega}_{tj} \cdot \mathrm{e}\right\|_2 \qquad \forall t \in \mathcal{T},
$$

where $\omega_{tj} : \mathcal{S}_j \times \mathcal{A}_j \to \mathbb{R}$ *are (arbitrary) weighting functions and* $\hat{\omega}_{tj} \in \mathbb{R}$ *are arbitrary constants,* $t \in \mathcal{T}$ *and* $j \in \mathcal{J}$*,* $\mathrm{e} : \mathcal{S}_j \times \mathcal{A}_j \to \mathbb{R}$ *is the all-one function satisfying* $\mathrm{e}(s,a) = 1$ *for all* $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$ *and the 2-norm is taken over the components* $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$ *of* $\omega_{tj} - \hat{\omega}_{tj} \cdot \mathrm{e}$.

Proof of Lemma 2. In view of the first statement, we note that

$$
\mathbb{E}\left[\tilde{\alpha}^{\star}_{tj}(s,a)\right] \;=\; \sum_{i=1}^{n_j} \mathbb{P}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right] \;=\; \sum_{i=1}^{n_j} \frac{\pi^{\mathrm{LP}}_{tj}(s,a)}{n_j} \;=\; \pi^{\mathrm{LP}}_{tj}(s,a),
$$

where the first and second identity follow from the definition of $\tilde{\alpha}^{\star}_{tj}(s,a)$ and Lemma 1, respectively.

As for the second statement, note that

$$\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\omega_{tj}(s,a)\cdot\left[\tilde{\alpha}^{\star}_{tj}(s,a)-\pi^{\mathrm{LP}}_{tj}(s,a)\right] \;=\; \sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\left[\omega_{tj}(s,a)-\hat{\omega}_{tj}\cdot\mathrm{e}\right]\cdot\left[\tilde{\alpha}^{\star}_{tj}(s,a)-\pi^{\mathrm{LP}}_{tj}(s,a)\right]$$

$\mathbb{P}$-a.s., where the identity holds because

$$\begin{aligned}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\left[\hat{\omega}_{tj}\cdot\mathrm{e}\right]\cdot\left[\tilde{\alpha}^{\star}_{tj}(s,a)-\pi^{\mathrm{LP}}_{tj}(s,a)\right] \;&=\; \sum_{j\in\mathcal{J}}\hat{\omega}_{tj}\cdot\left(\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\tilde{\alpha}^{\star}_{tj}(s,a)-\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\pi^{\mathrm{LP}}_{tj}(s,a)\right)\\ &=\; \sum_{j\in\mathcal{J}}\hat{\omega}_{tj}\cdot(n_j-n_j) \;=\; 0\end{aligned}$$

$\mathbb{P}$-a.s. Moreover, each $\tilde{\alpha}^{\star}_{tj}(s,a)$ is a sum of i.i.d. random variables; Hoeffding's inequality thus implies

$$\mathbb{P}\left[\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\left[\omega_{tj}(s,a)-\hat{\omega}_{tj}\right]\cdot\left[\tilde{\alpha}^{\star}_{tj}(s,a)-\pi^{\mathrm{LP}}_{tj}(s,a)\right]\leqslant\gamma\right] \;\geqslant\; 1-\exp\left(-\frac{2\gamma^2}{\sum_{j\in\mathcal{J}}n_j\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\left[\omega_{tj}(s,a)-\hat{\omega}_{tj}\right]^2}\right)$$

for all $t\in\mathcal{T}$ and $\gamma>0$. Since

$$\sum_{j\in\mathcal{J}}n_j\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j}\left[\omega_{tj}(s,a)-\hat{\omega}_{tj}\right]^2 \;\leqslant\; N\cdot\max_{j\in\mathcal{J}}\;\left\|\omega_{tj}-\hat{\omega}_{tj}\cdot\mathrm{e}\right\|_2,$$

the statement follows from the choice $\gamma=\sqrt{N\log N}\cdot\max_{j\in\mathcal{J}}\;\left\|\omega_{tj}-\hat{\omega}_{tj}\cdot\mathrm{e}\right\|_2$ and the union bound.

□

Proof of Theorem 1. In view of the bound on the expected total reward, we have

$$\begin{aligned}
\theta^\star &= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j} r_{jt}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)})\right] \\
&= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\mathbf{1}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right]\right] \\
&= \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\mathbb{E}\left[\tilde{\alpha}^\star_{tj}(s,a)\right] \\
&= \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} r_{jt}(s,a)\cdot\pi^{\mathrm{LP}}_{tj}(s,a) \\
&= \theta^{\mathrm{LP}} \;\geqslant\; \theta^{\mathrm{DP}},
\end{aligned}$$

where the first identity holds by definition of $\theta^\star$, the third identity follows from the definition of $\tilde{\alpha}^\star_{tj}(s,a)$ in Lemma 2, the fourth identity is due to Lemma 2, the last identity holds by definition of θ^{LP}, and the inequality holds since the fluid LP (4) is a relaxation of the weakly coupled DP.

As for the resource violation, with probability at least $1 - |\mathcal{T}|\,/\,N^2$ we have

$$\begin{aligned}
\sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) &= \sum_{j\in\mathcal{J}}\sum_{i=1}^{n_j}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\mathbf{1}\left[\tilde{s}_{t,(j,i)} = s \;\wedge\; \tilde{a}_{t,(j,i)} = a\right] \\
&= \sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\tilde{\alpha}^\star_{tj}(s,a) \\
&\leqslant \sum_{j\in\mathcal{J}}\sum_{s\in\mathcal{S}_j}\sum_{a\in\mathcal{A}_j} c_{tlj}(s,a)\cdot\pi^{\mathrm{LP}}_{tj}(s,a) \;+\; \sqrt{N\log N}\cdot\max_{j\in\mathcal{J}}\;\|c_{tlj}\|_2 \\
&\leqslant b_{tl} \;+\; \sqrt{N\log N}\cdot\max_{j\in\mathcal{J}}\;\|c_{tlj}\|_2
\end{aligned}$$

for all $t \in \mathcal{T}$ and any *fixed* $l \in \mathcal{L}$, where the second identity holds by definition of $\tilde{\alpha}^\star_{tj}(s,a)$ in Lemma 2, the first inequality follows from the second part of Lemma 2 if we choose $\omega_{tj} = c_{tlj}$ and $\hat{\omega}_{tj} = 0$, and the second inequality is implied by the third constraint of the fluid LP (4). The result then follows from an application of the union bound to the $|\mathcal{L}|$ resource constraints $l \in \mathcal{L}$. □

Proof of Theorem 2. The bound on the resource violation is the same as in Theorem 1, and we refer to its proof for the justification of the bound. In view of the bound on the worst-case total reward, we observe that with probability at least $1 - |\mathcal{T}| \, / \, N^2$, we have

$$\begin{aligned}
\tilde{\theta}^\star \;&=\; \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} r_{jt}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \\
&=\; \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a) \cdot \tilde{\alpha}^\star_{tj}(s,a) \\
&\geqslant\; \sum_{t \in \mathcal{T}} \left[\sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_j} \sum_{a \in \mathcal{A}_j} r_{jt}(s,a) \cdot \pi^{\mathrm{LP}}_{tj}(s,a) \;-\; \sqrt{N \log N} \cdot \max_{j \in \mathcal{J}} \; \| r_{jt} \|_2 \right] \\
&=\; \theta^{\mathrm{LP}} \;-\; \sum_{t \in \mathcal{T}} \sqrt{N \log N} \cdot \max_{j \in \mathcal{J}} \; \| r_{jt} \|_2 \\
&\geqslant\; \theta^{\mathrm{DP}} \;-\; |\mathcal{T}| \cdot \sqrt{N \log N} \cdot \max_{\substack{t \in \mathcal{T}, \\ j \in \mathcal{J}}} \; \| r_{jt} \|_2 \, ,
\end{aligned}$$

where the first identity is due to the definition of $\tilde{\theta}^\star$, the second identity follows from the definition of $\tilde{\alpha}^\star_{tj}(s,a)$ in Lemma 2, the first inequality follows from the second part of Lemma 2 if we choose $\omega_{tj} = -r_{jt}$ and $\hat{\omega}_{tj} = 0$, the last identity follows from the definition of θ^{LP}, and the second inequality holds since the fluid LP (4) is a relaxation of the weakly coupled DP. The result then follows if we apply the union bound to jointly bound *(i)* the probability of the total reward violating the above inequality and *(ii)* any of the resource constraints being violated. □

Proof of Proposition 3. We prove the statement in two steps. We first show that the optimal value of the fluid LP (4) does not decrease significantly if we reduce the resource limits from $\{b_{tl}\}_{t,l}$ to $\{b^{\mathrm{red}}_{tl}\}_{t,l}$. Note that the expected total reward of any optimal policy for the grouped weakly coupled DP is bounded above by the optimal value of the fluid LP (4) under the original resource limits, and that the proof of Theorem 1 implies that the expected total reward of the auxiliary policy π^{red} equals the optimal value of the fluid LP (4) under the updated resource limits. We then argue in a second step that the feasible randomized policy π^{F} deviates from the auxiliary policy π^{red} only with small probability, and that the impact of those deviations can be bounded.

In view of the first step, denote by $\theta^{\mathrm{LP}} : \mathbb{R}^{LT}_{+} \rightarrow \mathbb{R}$ the function that maps resource limits to the optimal value of the corresponding fluid LP (4). Standard results from linear programming imply that θ^{LP} is non-decreasing and concave. We thus observe that

$$\theta^{\mathrm{LP}}(b^{\mathrm{red}}) \;\geqslant\; \theta^{\mathrm{LP}}\left(\min_{\substack{t\in\mathcal{T},\\ l\in\mathcal{L}}}\left\{\frac{b_{tl}^{\mathrm{red}}}{b_{tl}}\right\}\cdot b\right) \;\geqslant\; \min_{\substack{t\in\mathcal{T},\\ l\in\mathcal{L}}}\left\{\frac{b_{tl}^{\mathrm{red}}}{b_{tl}}\right\}\cdot\theta^{\mathrm{LP}}(b) \;+\; \left(1-\min_{\substack{t\in\mathcal{T},\\ l\in\mathcal{L}}}\left\{\frac{b_{tl}^{\mathrm{red}}}{b_{tl}}\right\}\right)\cdot\theta^{\mathrm{LP}}(0), \tag{8}$$

where $b = \{b_{tl}\}_{t,l}$ and $b^{\mathrm{red}} = \{b_{tl}^{\mathrm{red}}\}_{t,l}$ represent the original and updated resource limits of the grouped weakly coupled DP, respectively. Here, the first inequality follows from the fact that θ^{LP} is non-decreasing, while the second equality is due to the concavity of θ^{LP}. Since

$$\begin{aligned}
\min_{\substack{t\in\mathcal{T},\\ l\in\mathcal{L}}}\left\{\frac{b_{tl}^{\mathrm{red}}}{b_{tl}}\right\} \;&=\; \min_{\substack{t\in\mathcal{T},\\ l\in\mathcal{L}}}\left\{\left[b_{tl}-\sqrt{N\log N}\cdot\max_{j\in\mathcal{J}}\,\|c_{tlj}\|_2\right]_+ \Big/ b_{tl}\right\}\\
&\geqslant\; \min_{\substack{t\in\mathcal{T},\\ l\in\mathcal{L}}}\left\{\left(b_{tl}-\sqrt{N\log N}\cdot\max_{j\in\mathcal{J}}\,\|c_{tlj}\|_2\right)\Big/ b_{tl}\right\}\\
&=\; 1-\sqrt{\frac{\log N}{N}}\cdot\max_{\substack{t\in\mathcal{T},\,l\in\mathcal{L},\\ j\in\mathcal{J}}}\left\{\frac{N}{b_{tl}}\cdot\|c_{tlj}\|_2\right\}
\end{aligned}$$

and $\theta^{\mathrm{LP}}(0) \geqslant N\cdot|\mathcal{T}|\cdot\underline{r}$ by construction, we thus obtain that

$$\begin{aligned}
\theta^{\mathrm{LP}}(b^{\mathrm{red}}) \;&\geqslant\; \min_{\substack{t\in\mathcal{T},\\ l\in\mathcal{L}}}\left\{\frac{b_{tl}^{\mathrm{red}}}{b_{tl}}\right\}\cdot\left[\theta^{\mathrm{LP}}(b)-\theta^{\mathrm{LP}}(0)\right]+\theta^{\mathrm{LP}}(0)\\
&\geqslant\; \left(1-\sqrt{\frac{\log N}{N}}\cdot\max_{\substack{t\in\mathcal{T},\,l\in\mathcal{L},\\ j\in\mathcal{J}}}\left\{\frac{N}{b_{tl}}\cdot\|c_{tlj}\|_2\right\}\right)\cdot\left[\theta^{\mathrm{LP}}(b)-\theta^{\mathrm{LP}}(0)\right]+\theta^{\mathrm{LP}}(0)\\
&\geqslant\; \left(1-\sqrt{\frac{\log N}{N}}\cdot\max_{\substack{t\in\mathcal{T},\,l\in\mathcal{L},\\ j\in\mathcal{J}}}\left\{\frac{N}{b_{tl}}\cdot\|c_{tlj}\|_2\right\}\right)\cdot\theta^{\mathrm{LP}}(b) &(9)\\
&\qquad +\sqrt{\frac{\log N}{N}}\cdot\max_{\substack{t\in\mathcal{T},\,l\in\mathcal{L},\\ j\in\mathcal{J}}}\left\{\frac{N}{b_{tl}}\cdot\|c_{tlj}\|_2\right\}\cdot N\cdot|\mathcal{T}|\cdot\underline{r}, &(10)
\end{aligned}$$

where the first inequality holds because of (8) and the second one holds since θ^{LP} is non-decreasing.

As for the second step, we note that the auxiliary policy π^{red} (which was constructed under the updated resource limits b^{red}) violates the original resource limits b with probability at most $|\mathcal{T}|\cdot|\mathcal{L}|/N^2$. Indeed, for $(t,l)\in\mathcal{T}\times\mathcal{L}$ with $b_{tl}^{\mathrm{red}} > 0$, this follows from Theorem 1 and the construction of b^{red}, whereas for $(t,l)\in\mathcal{T}\times\mathcal{L}$ with $b_{tl}^{\mathrm{red}} = 0$, this holds due to the fact that the optimal solution $(\sigma^{\mathrm{LP}},\pi^{\mathrm{LP}})$ to the fluid LP (4) with resource limits b^{red} will satisfy $\pi_{tj}^{\mathrm{LP}}(s,a) = 0$

for all $(s,a) \in \mathcal{S}_j \times \mathcal{A}_j$ where $c_{tlj}(s,a) > 0$. The feasible randomized policy π^{F} coincides with the auxiliary policy π^{red} on state-action trajectories where π^{red} satisfies all of the original resource limits, whereas the actually realized total reward of π^{F} falls below that of π^{red} by at most $N \cdot |\mathcal{T}| \cdot (\overline{r} - \underline{r})$ on state-action trajectories where π^{red} violates some of the original resource limits. Thus, the expected total rewards θ^{F} and θ^{red} of the policies π^{F} and π^{red}, respectively, satisfy

$$\theta^{\mathrm{F}} \;\geqslant\; \theta^{\mathrm{red}} - \frac{|\mathcal{T}| \cdot |\mathcal{L}|}{N^2} \cdot [N \cdot |\mathcal{T}| \cdot (\overline{r} - \underline{r})] \;=\; \theta^{\mathrm{red}} - \frac{|\mathcal{T}|^2 \cdot |\mathcal{L}| \cdot (\overline{r} - \underline{r})}{N}.$$

Recalling from the proof of Theorem 1 that $\theta^{\mathrm{red}} = \theta^{\mathrm{LP}}(b^{\mathrm{red}})$ and using (9), we thus obtain

$$\begin{aligned}\theta^{\mathrm{F}} \;&\geqslant\; \theta^{\mathrm{red}} - \frac{|\mathcal{T}|^2 \cdot |\mathcal{L}| \cdot (\overline{r} - \underline{r})}{N} \;=\; \theta^{\mathrm{LP}}(b^{\mathrm{red}}) - \frac{|\mathcal{T}|^2 \cdot |\mathcal{L}| \cdot (\overline{r} - \underline{r})}{N} \\ &\geqslant\; \left(1 - \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T},\, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \cdot \|c_{tlj}\|_2 \right\}\right) \cdot \theta^{\mathrm{LP}}(b) \\ &\quad + \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T},\, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \cdot \|c_{tlj}\|_2 \right\} \cdot N \cdot |\mathcal{T}| \cdot \underline{r} - \frac{|\mathcal{T}|^2 \cdot |\mathcal{L}| \cdot (\overline{r} - \underline{r})}{N}.\end{aligned}$$

The statement now follows from the fact that $\theta^{\mathrm{LP}}(b) \geqslant \theta^{\mathrm{DP}}$. □

Proof of Proposition 4 Denote by $\tilde{\theta}^{\mathrm{red}}$ the random total reward of the auxiliary policy π^{red} constructed from an optimal solution to the fluid LP (4) with the updated resource limits b^{red}. According to the statement and proof of Theorem 2, we have

$$\tilde{\theta}^{\mathrm{red}} \;\geqslant\; \theta^{\mathrm{LP}}(b^{\mathrm{red}}) \;-\; |\mathcal{T}| \cdot \sqrt{N \log N} \cdot \max_{\substack{t \in \mathcal{T}, \\ j \in \mathcal{J}}} \|r_{jt}\|_2 \tag{11a}$$

as well as

$$\sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \;\leqslant\; b_{tl}^{\mathrm{red}} + \sqrt{N \log N} \cdot \max_{j \in \mathcal{J}} \|c_{tlj}\|_2 \qquad \forall t \in \mathcal{T},\ \forall l \in \mathcal{L}, \tag{11b}$$

both jointly with probability at least $1 - |\mathcal{T}| \cdot (|\mathcal{L}| + 1)/N^2$. The inequality (11a) implies that with probability at least $1 - |\mathcal{T}| \cdot (|\mathcal{L}| + 1)/N^2$, we have

$$\begin{aligned}
\tilde{\theta}^{\mathrm{red}} \;&\geqslant\; \theta^{\mathrm{LP}}(b^{\mathrm{red}}) \;-\; |\mathcal{T}| \cdot \sqrt{N \log N} \cdot \max_{\substack{t \in \mathcal{T}, \\ j \in \mathcal{J}}} \|r_{jt}\|_2 \\
&\geqslant\; \left(1 - \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T}, \, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \cdot \|c_{tlj}\|_2 \right\} \right) \cdot \theta^{\mathrm{LP}}(b) \\
&\quad + \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T}, \, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \cdot \|c_{tlj}\|_2 \right\} \cdot N \cdot |\mathcal{T}| \cdot \underline{r} \;-\; |\mathcal{T}| \cdot \sqrt{N \log N} \cdot \max_{\substack{t \in \mathcal{T}, \\ j \in \mathcal{J}}} \|r_{jt}\|_2 \\
&\geqslant\; \left(1 - \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T}, \, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \cdot \|c_{tlj}\|_2 \right\} \right) \cdot \theta^{\mathrm{DP}} \\
&\quad + \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T}, \, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \cdot \|c_{tlj}\|_2 \right\} \cdot N \cdot |\mathcal{T}| \cdot \underline{r} \;-\; |\mathcal{T}| \cdot \sqrt{N \log N} \cdot \max_{\substack{t \in \mathcal{T}, \\ j \in \mathcal{J}}} \|r_{jt}\|_2 \,,
\end{aligned}$$

where the second inequality is due to the proof of Proposition 3, and the third inequality holds since the fluid LP (4) constitutes a relaxation of the grouped weakly coupled DP. As for the inequality (11b), on the other hand, the definition of the updated resource limits b^{red} implies that

$$b^{\mathrm{red}}_{tl} + \sqrt{\frac{\log N}{N}} \cdot \max_{\substack{t \in \mathcal{T}, \, l \in \mathcal{L}, \\ j \in \mathcal{J}}} \left\{ \frac{N}{b_{tl}} \cdot \|c_{tlj}\|_2 \right\} \;=\; b_{tl}$$

for all $(t, l) \in \mathcal{T} \times \mathcal{L}$ satisfying $b^{\mathrm{red}}_{tl} > 0$, while

$$\sum_{j \in \mathcal{J}} \sum_{i=1}^{n_j} c_{tlj}(\tilde{s}_{t,(j,i)}, \tilde{a}_{t,(j,i)}) \;=\; 0 \;\leqslant\; b_{tl}$$

for all $(t, l) \in \mathcal{T} \times \mathcal{L}$ satisfying $b^{\mathrm{red}}_{tl} = 0$ (*cf.* the proof of Proposition 3). In other words, the auxiliary policy π^{red} adheres to the original resource limits b with high probability. The statement of the proposition now follows from the fact that the feasible randomized policy π^{F} coincides with the auxiliary policy π^{red} whenever the latter satisfies the original resource limits, which according to the above discussion happens with high probability. □