manuscript

The Analytics of Bed Shortages:

Coherent Metric, Prediction and Optimization

Jingui Xie
School of Management, University of Science and Technology of China, Hefei, China 230026, xiej@ustc.edu.cn

Gar Goei Loke
Department of Mathematics, National University of Singapore, Singapore 119076, e0012863@u.nus.edu

Melvyn Sim
NUS Business School, National University of Singapore, Singapore 119245, dscsimm@nus.edu.sg

Shao Wei Lam
Health Services Research Centre, Singapore Health Services, Singapore 169856, lam.shao.wei@singhealth.com.sg

Bed shortages in hospitals usually have a negative impact on patient satisfaction and medical outcomes. In practice, healthcare managers often use bed occupancy rates (BOR) as a metric to understand bed utilization, which is insufficient in capturing the risk of bed shortages. We propose the *bed shortage index* (BSI) to capture more facets of bed shortage risk than traditional metrics such as the occupancy rate, the probability of shortages and expected shortages. The BSI is based on the well-known Aumann and Serrano (2008) riskiness index and it is calibrated to coincide with BOR when the daily arrivals in the hospital unit are Poisson distributed. Our metric can be tractably computed and does not require additional assumptions or approximations. As such, it can be consistently used across the descriptive, predictive and prescriptive analytical approaches. We also propose optimization models to plan for bed capacity via this metric. These models can be efficiently solved on a large scale via a sequence of linear optimization problems. The first maximizes total elective throughput while managing the metric under a specified threshold. The second determines the optimal scheduling policy by lexicographically minimizing the steady-state daily BSI for a given number of scheduled admissions. We validate these models using real data from a hospital and test them against data-driven simulations. We apply these models to study the real-world problem of long stayers, to predict the impact of transferring them to community hospitals, as a result of an aging population.

Key words: Analytics, bed shortages, coherent metric, riskiness index, data-driven optimization, simulation

1. Introduction

Bed shortages can compromise the ability of hospitals to achieve desirable patient outcomes and maintain service quality. A variety of factors may contribute to bed shortages. In particular, it can be driven by changes in demand for beds, such as a rise in admissions and longer lengths of stay (LOS). For example, an ageing population is one such major driving factor. In World Health Organisation's publication of World Health Statistics 2016, 29 countries, in total accounting for

at least 650 million of the global population, now have an average life expectancy of 80 years and older. With better healthcare systems, the average life expectancy would only be on an upward trend. For Singapore in particular, the average life expectancy for women has risen from 66 in 1960 to 84 years, and from 62 to 79 years for men (Haseltine 2013). Correspondingly, the healthcare burden is expected to increase in every aspect – manpower and cost of caregivers, and of course, the risk of bed shortages in hospitals. Fluctuations in the patterns of demand can also lead to bed shortages. Increasingly, with urbanization, there is a higher threat of demand surges. Incidences of terrorism, pandemics and natural disasters now have a larger potential than ever to create mass casualty events (Challen et al. 2007).

These trends and challenges have made it ever more important to better manage beds and mitigate the risk of shortages. The following might constitute a suite of bed management problems:

1. How many beds does a hospital need in order to serve its immediate population?
2. How much capacity does a hospital need to deal effectively with demand surges?
3. How should a hospital system planner demarcate the zones for ambulances?
4. How should transfer policies between hospitals of different specializations be fixed, especially the transfer of patients out of an acute care hospital to one specializing in long-term care?

These are challenging questions. For example, in 2010, a new 550-bed hospital was opened in Singapore. It, however, had little impact on easing bed shortages across the nation-state[1]. Utilization is also growing – the average LOS of patients in Singapore's public hospitals has crept up from 7.8 days in 2010 to 8.2 in 2013[2].

The literature holds an abundance of approaches focusing on the matter of adaptive scheduling. From the angle of healthcare planners, however, long-term bed capacity planning is equally important. In this paper, we are interested in models to jointly plan for capacity and elective scheduling; we refer to these as 'steady-state' problems. To address problems like the four aforementioned ones, we hope to create a decision-making framework that is general enough for most capacity planning and scheduling problems and yet specific enough to model each problem context.

In the literature, queueing theory and stochastic simulations (*e.g.*, Lamiri et al. 2008, Cochran and Roche 2008) are popularly employed. The benefit of the queueing paradigm is that it relates traditional metrics, such as the Bed Occupancy Rates (BOR) to the probability of shortages and expected delays, from which planners can decide on the capacity and their responses. The queueing approach is also especially popular in studying the inpatient flow problem (as in Shi et al. 2016). For example, Dai and Shi (2018) model hospital inpatient flow as a multi-class queueing system and formulate the overflow decision problem as a discrete-time Markov decision process. They study the trade-off between overflow and congestion using approximate dynamic programming.

Nonetheless, these models may not be adequate for addressing the adaptive bed management problem. Critically, the full complexity associated with the LOS of patients can be hard to capture. Additionally, the queueing paradigm may depart from the actual dynamics of patient admissions. Recent empirical studies show that when beds are fully occupied, new arrivals are instead treated in other non-primary wards (Shi et al. 2016, Song et al. 2018). This is termed 'patient overflow'. Helm and Van Oyen (2014) illustrate its extent – in their data, around 17% patients were accommodated and treated in overflow wards. They subsequently apply an infinite capacity *offered load* model, which continues to render service to patients using "temporary" beds instead of keeping the blocked patients waiting. They argue that the offered load model may be a better representation of reality than a loss model or a traditional queueing model. In our paper, we adopt the same assumptions on patient dynamics, but compare our results with a traditional queueing model where capacity is fixed and inflexible.

Stochastic and robust optimization models have also emerged to address these problems (*e.g.*, Meng et al. 2015, Qi 2017). Gupta and Denton (2008) provide an excellent review of appointment scheduling systems which are used by clinics as well as hospitals to schedule elective surgeries. We draw particular point of note to Helm and Van Oyen (2014). They develop linearizing approximations to stochastic metrics capturing patient blocking and incorporate them into a mixed-integer linear optimization problem (MILP) to solve the hospital admission scheduling and control problem.

At present, models in the literature are unable to achieve a high level of complexity. To obtain tractable solutions, simplifying assumptions are made, often also accompanied by context-specific linearizations. This is especially apparent when traditional metrics of bed shortages are adopted, such as the probability of shortages, or expected shortage. For example, Helm and Van Oyen (2014) assume that the number of elective patients in a hospital unit on each day is deterministic in order to arrive at a tractable problem. Meng et al. (2015) avert this issue in their formulation by ignoring important statistical information such as independence. These approaches fundamentally identify a gap in the literature – tractability assumptions and linearizations are inescapable. Hence, models which scale to address longer-term planning issues, while accurately capturing the risks and dynamics at the local and ward-level, remain elusive. In addition, the need to incorporate increasingly proliferated predictive analyses (such as in Lynn et al. 2007, Zhou et al. 2014) into an optimization model for capacity planning is growing. We hope to build a model that has this flexibility.

To do so, we first seek a metric that captures the risk of bed shortages and which is amenable to optimization. This metric can then be used in conjunction with other traditional metrics in practice. In the paper, we describe a *coherent bed shortage metric* and characterize its representation. In

particular, we propose the *Bed Shortage Index* (BSI), built upon the riskiness index of Aumann and Serrano (2008). The BSI, which can be considered as a risk-adjusted form of the BOR, is a functional that maps the distribution of bed shortages to a number between zero and one, with zero equating to no risk of bed shortages and one corresponding to high risks associated with overloading. We show that if all arrivals are Poisson, the BSI coincides with the BOR. If there is greater uncertainty, such as in the arrival rates, then the BSI would capture the risk and hence be larger than the BOR. Conversely, if all arrivals are certain and scheduled, then the BSI is smaller than the BOR, indicating lower risk of shortages.

The BSI possesses the following advantages:

1. It is easily interpreted by healthcare practitioners and administrators. The BSI of a hospital unit is the equivalent level of bed shortage risk experienced by a referenced unit that is serving only Poisson arrivals, with its BOR being the same as BSI (Theorem 2).
2. It has an exact closed-form reformulation (Theorem 4), thus circumventing the need for context-specific approximations, and/or tractability assumptions. It thus can be flexibly applied to different problem contexts, while preserving its fidelity to the actual dynamics of patient flows.
3. It captures higher-order statistical information compared to other traditional metrics, and hence is a sharper articulation of the risks of bed shortages.
4. It can be decomposed at the level of patient segmentations and hence is a valid tool for strategic-level future capacity planning that incorporates predictive analyses (Appendix C).
5. It relates to the traditional metrics of probability of shortages and expected shortages (Proposition 2).

The contributions of this paper are summarized as follows. First, our paper considers a general problem of capacity planning and resource allocation in hospitals. This departs from the usual approach in the literature, which is concerned largely with the problem of adaptive bed management in the short-term. In this vein, our paper seeks not to solve a particular instance of the bed management problem, but to provide a framework to a whole suite of them, ranging from capacity planning at the ward level, to contingency planning for demand surges, to elective scheduling, and to tandem planning between acute care and community hospitals. Indeed, we describe technically or numerically how our model can be applied to all of these problems to illustrate the sheer generalizing quality of our model. As such, the proposed approach needs to be flexible enough to model different contexts and yet accurately model the inherent dynamics in the problem. In this regard, we cannot understate the importance of Theorem 4. The fact that exact reformulations of the optimization problem are recovered, means that our framework does not need to concern

itself with problem-specific linearizations or approximations. To the limits of our knowledge, we are unable to find another model in the literature that is as general or flexible.

Second, to the best of our knowledge, our model is the first model that attempts to harmonize different approaches of analytics, by drawing the connection between the wealth of analyses generated at the predictive level and strategic-level decision-making. We believe this to be an increasingly important paradigm in the age of analytics. Decisions on capacity planning have to take into consideration a variety of factors, in particular, an accurate representation of the risk, an understanding of how this risk shifts under changing circumstances, and a means of synthesizing these analyses in order to optimize strategically capacity planning decisions. This cannot be reasonably achieved with present-day bed management models. Our model can handle this optimization easily, precisely also because of the flexibility accorded by the result of Theorem 4.

Notation

Unless otherwise specified, all variables are known constants. Random or uncertain variables are accented with the tilde, for example, $\tilde{z} \sim \mathbb{P}$. We use the convention, $\sup \emptyset = -\infty$ and $\inf \emptyset = \infty$ where $\emptyset$ is the empty set. We define $x^+ \triangleq \max\{0, x\}$. We say that a functional $\rho : \mathcal{Z} \to \mathbb{R}$ mapping out of a set of random variables $\mathcal{Z}$ defined on probability space $(\Omega, \Sigma, \mathbb{P})$ is *lower semi-continuous* if and only if the set $\{\tilde{z} \mid \rho(\tilde{z}) \leq a\}$ is closed for all $a \in \mathbb{R}$.

2. Coherent metrics for evaluating bed shortages

Often, healthcare practitioners study shortages via the proxy of *bed occupancy rates* (BOR). This is defined as the ratio of occupied beds against the total number of beds in service. For many years, the BOR has served practitioners as the primary measure guiding bed capacity decisions, and the common wisdom is to keep the BOR under 85% (Green 2002). In Singapore, for example, BORs have been fluctuating above 85% for most public hospitals. Researchers have also related BOR to various patient outcomes. High BOR can lead to high risk of bed shortages, which has been shown to result in prolonged LOS in Emergency Departments (ED) (*e.g.*, Schull et al. 2001, Forster et al. 2003) and increased risk of hospital acquired infections (*e.g.*, Kaier et al. 2012, Clements et al. 2008). The association between mortality rates and high BOR has also been reported in recent studies (*e.g.*, Sprivulis et al. 2006, Madsen et al. 2014). A high BOR has also been shown to have an impact on providers' satisfaction and morale (Virtanen et al. 2008). As a high BOR reduces the resilience of hospitals in responding to uncertainties, balancing between emergency and elective inpatient admissions is an active problem for healthcare administrators (Meng et al. 2015). In response, various strategies have been proposed. On the demand side, scheduling elective admissions while taking into account the predicted volume of emergency admissions and discharges

is a common strategy (*e.g.*, Meng et al. 2015, Shi et al. 2016), whereas on the supply side, dynamic bed overflow policies have also been considered (Teow et al. 2012, Dai and Shi 2018).

While the BOR may be easily understood by practitioners and tractably computed, it is not an adequate metric for articulating the risk of bed shortages (see discussions in Green 2002). We also illustrate later that using the BOR as a metric for optimization can lead to sub-optimal outcomes.

Often the BOR is also used in conjunction with other metrics when planning for bed utilization, including the probability of shortages (or delays), the expected shortages (or overflows as in Kao and Tung 1981) and expected delays. Indirect measures, such as the costs of shortages and holding (Esogbue and Singh 1976), or one which incorporates economic efficiency explicitly (such as in Kuntz et al. 2007, where the metric was illustrated within a hospital capacity planning cycle), have been proposed.

Regardless, these metrics have some drawbacks. Most critically, these measures may not be readily amenable for optimization under uncertainty, for example, the probability of shortages and expected shortages. Instead, further assumptions or approximations would have to adopted (for example, linearizing the Erlang loss formula for expected shortage, as in Helm and Van Oyen 2014). Some of these assumptions, however as discussed in the Introduction, may not be most natural in capturing the dynamics of patient flows.

It can also be difficult to capture higher-order statistical information with these metrics. This leads many stochastic approaches to utilize only known distributional characteristics. We later illustrate the sub-optimality of making decisions based on statistics with a lower degree of fidelity.

Apart from that, some of these metrics may not be readily observable, for example, the cost of shortage and holding. Decision-making also predicates a risk appetite which the administrator must articulate. This is rarely explicit in the decision support frameworks utilizing these metrics.

In practice, these metrics are seldom used in isolation to make decisions regarding the bed capacity of wards. Nonetheless, prioritizing a metric over another comes down to a modelling choice, which can affect the optimal policy in ways not necessarily known at the onset. As such, it may not necessarily be reasonable to expect the practitioner to decide on it *a priori*, and hence there may be a natural deference to adopt whichever metric has been the set norm.

We attempt to circumvent these problems. We desire a metric that captures the risks of bed shortages by incorporating higher-order statistical information, without compromising its computational tractability. It needs to embody the ease of interpretation so that health practitioners and administrators could comprehend and specify its parameters, like the BOR. This enables the metric to be seamlessly integrated through the entire value stream of analytics, from descriptive statistics to predictive analytics, and into prescriptive optimization models. We embark on this below.

2.1. Bed shortages

Let T be the planning time horizon and $\mathcal{T} = \{1, \ldots, T\}$. To capture variations in bed shortages over the week, we use $T = 7$ in this paper, with Days 1 to 7 corresponding to Monday through Sunday. A planner who is indifferent to daily variations may let $T = 1$ instead. On a particular day $t \in \mathcal{T}$, let $\tilde{n}_t$ be the random variable that represents the demand for beds. This demand may be the sum of new admissions and the possibility that each present inpatient remains warded for an additional day. Denote by κ_t the ward capacity on day t.

Definition 1 (bed shortage position). We define the *bed shortage position* on day t as the random variable $\tilde{z}_t \triangleq \tilde{n}_t - \kappa_t$. Hence, $\tilde{z}_t^+ \triangleq \max\{0, \tilde{z}_t\}$ denotes *bed shortages*.

There is often a difference in the patterns between emergency and elective patients, as illustrated numerically later. We make this explicit. Define $\bar{\mathcal{T}} = \{t \in \mathbb{Z} \mid t \leq T\}$ as the set of times up to T, including the past. For $t \in \bar{\mathcal{T}}$ and $s \geq 0$, define $\tilde{x}_{t,s}$ as the number of elective patients admitted on day t and who has remained in the ward at the end of day $t+s$ (*i.e.*, the patient has spent at least $s+1$ days in the ward). Similarly, define $\tilde{y}_{t,s}$ as the number of emergency patients who were admitted on day t and has remained in the ward at the end of day $t+s$. Let η_t be the scheduled quota of elective patients on day t.

Uncertainty is assumed in the definition of $\tilde{x}_{t,s}$ and $\tilde{y}_{t,s}$ – emergency patients arrive randomly and both elective and emergency patients have a random LOS in the ward. Hence, for each $s \geq 0$, denote by p_s the probability that a scheduled elective patient stays in the hospital at least until the end of the $(s+1)$-th day (*i.e.*, the survival rate for $s+1$ days). Although quotas are usually filled in advance, patients may not turn up or be warded for at least a day. In this case, $1 - p_0$ gives the no-show probability. Likewise, if λ_t denotes the arrival rate of emergency patients on day t, then q_s is the probability that the emergency patient stays in the hospital at least until the end of the $(s+1)$-th day. Since by definition every admitted emergency patient would stay for at least 1 day, we have $q_0 = 1$. In this paper, we assume that:

Assumption 1 **(Model of Uncertainties)**.

(a) The stochastic arrivals of all patients and their lengths of stay are independent.

(b) Emergency arrivals follow a non-homogeneous Poisson process with random rates: $\tilde{y}_{t,0} \sim Pois(\tilde{\lambda}_t)$, where the log moment-generating function of $\tilde{\lambda}_t$ is $\Lambda_t(\theta) \triangleq \log \mathbb{E}[\exp(\tilde{\lambda}_t \theta)]$. If $\tilde{\lambda}_t = \lambda_t$ is a constant, then $\Lambda_t(\theta) = \lambda_t \theta$.

Assumption 1 not only includes the classical assumption of a non-homogeneous Poisson process with known but time-varying rates, but also extends to a more general assumption of a doubly stochastic Poisson process where the arrival rates are random (Whitt 1999, Avramidis et al. 2004).

This gives a better fit with essential Poisson and non-Poisson properties of arrivals observed in recent empirical studies (Kim and Whitt 2014, Kim et al. 2015). Despite the generality of Assumption 1, the actual dynamics could potentially be even more complicated. For example, LOS may depend on the number of people in the ward or delays experienced by patients awaiting care. As such, the two uncertainties may be correlated. KC and Terwiesch (2009, 2012) and Anderson et al. (2011) find that a high BOR can result in shorter patient LOS as it incentivizes early discharges to accommodate more critical patients. Chan et al. (2016) argue that long delays may have adverse effects on patient outcomes and can potentially lead to a longer LOS.

In our model, we do not wish to impose further assumptions on LOS. For computation simplicity, we define L as the maximum LOS so that $p_s = 0$ and $q_s = 0$ for all $s \geq L$. Under Assumption 1, we can show that the number of patients at day t, $\tilde{n}_t$, is a summation of elementary random variables.

PROPOSITION 1. *Under Assumption 1, the number of patients at day t, $\tilde{n}_t$, is a sum of Poisson binomial random variables for elective patients and Poisson random variables with random rates for emergency arrivals. More specifically,*

$$\tilde{n}_t = \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} (\tilde{x}_{\tau,s} + \tilde{y}_{\tau,s}),$$

where $\tilde{x}_{\tau,s} \sim Bin(\eta_\tau, p_s)$ and $\tilde{y}_{\tau,s} \sim Pois(\tilde{\lambda}_\tau q_s)$.

Proof. The proof is an easy consequence of Assumption 1. □

REMARK 1. By including an index i as we do later in Section 4, Proposition 1 and subsequent results including Theorem 4 can be easily extended to multiple classes of patients, *e.g.*, according to disease type and severity, patient age and gender, or clinical pathway, and multiple wards.

2.2. Coherent bed shortage metrics

To adequately take action on bed shortages, the planner would be interested in quantifying the risky bed shortage positions by providing a value that relates to the likelihood and severity of potential shortages. One option is to consider the BOR, defined as $\mathbb{E}[\tilde{n}_t/\kappa_t]$. Ostensibly, as the BOR approaches 100%, the bed shortage risk would also be greater. One of the main reasons for the popularity of BOR as a proxy to the risk of bed shortages is its simplicity. Indeed, to evaluate the BOR, we only require the expected daily demand of beds, $\tilde{n}_t$. However, because the BOR ignores pertinent distributional information of $\tilde{n}_t$, it is not an effective proxy. Hence, we have to construct a new metric for this purpose.

Let $(\Omega, \Sigma, \mathbb{P})$ be some probability space, where Ω is countable. Let $\mathcal{Z}$ be the set of discrete random variables on this probability space. We seek a metric $\beta : \mathcal{Z} \to [0,1]$ quantifying the bed shortage risk of a ward on a particular day, and behaves like the BOR, *i.e.* $\beta(\tilde{z}) \in [0,1]$ where $\beta(\tilde{z}) = 1$ represents an undesirable situation that the ward must avoid at all costs.

More specifically, we desire the following properties of β:

Definition 2 (Coherent Bed Shortage Metric). The functional $\beta : \mathcal{Z} \to [0,1]$ is a *coherent bed shortage metric* if and only if β is lower semi-continuous and for all bed shortage positions $\tilde{z}, \tilde{z}_1, \tilde{z}_2 \in \mathcal{Z}$ it satisfies:

(a) **Monotonicity**: If $\mathbb{P}[\tilde{z}_1 \leq \tilde{z}_2] = 1$, then $\beta(\tilde{z}_1) \leq \beta(\tilde{z}_2)$.

(b) **Risk-free**: If $\mathbb{P}[\tilde{z} \leq 0] = 1$, then $\beta(\tilde{z}) = 0$.

(c) **Overloading**: If $\mathbb{E}[\tilde{z}] > 0$, then $\beta(\tilde{z}) \geq 1$.

(d) **Quasiconvexity**: For all $\lambda \in [0,1]$, $\beta(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2) \leq \max\{\beta(\tilde{z}_1), \beta(\tilde{z}_2)\}$.

Additionally, we require β to be **Tractable**: It must be possible to compute β and optimize over it efficiently, and **Interpretable**: It should be understandable to healthcare practitioners and administrators.

These properties here are designed to behave akin to how the usual BOR, probability of shortages, or expected shortages would. The monotonicity property enforces that if almost surely fewer bed shortages occur, then indeed fewer beds should be occupied, *i.e.*, the occupancy rate is lower. The risk-free property guarantees that if bed shortages will emerge under no circumstance, then in effect, there is no risk to beds being occupied. This situation may occur if the ward is under-utilized. Instead, if shortages will emerge in expectation, then the ward is completely occupied with no spare capacity, *i.e.*, the coherent bed shortage metric hits 100% consistently, which indicates overloading. Hence, in a well managed ward, the coherent bed shortage metric should be between 0 and 1. In economics, quasiconvexity is synonymous with *convex preference*. Informally, this is the belief that 'averages are better than the extremes', which is associated with risk aversion. This is often assumed in the context of rational decision-making (Andreoni and Miller 2002). From the prescriptive perspective, this property induces a preference against extreme bed shortages, which may impact the mortality of the patients. For instance, the probability of shortages, a metric that is indifferent to the magnitudes of shortages, would violate quasiconvexity. Another important consideration for such a metric is its tractability, *i.e.*, amenability to computation and optimization. More formally, suppose the bed shortage position, $\tilde{z}(\boldsymbol{w})$ is a functional of a decision vector, $\boldsymbol{w} \in \mathbb{R}^n$. Then the separation problem associated with the level set $\{\boldsymbol{w} \mid \beta(\tilde{z}(\boldsymbol{w})) \leq \epsilon\}$ can be determined in polynomial time. In fact, as discussed earlier, approaches in the literature utilizing the traditional metrics remain fraught with tractability challenges.

More broadly, let $u:\mathbb{R}\to\mathbb{R}$ be a non-decreasing disutility function. The *expected disutility* metric, $\mathbb{E}[u(\tilde{z})]$, has been ubiquitous for evaluating risky positions and portfolio optimization. If u is increasing as well, then $u^{-1}(\mathbb{E}[u(\tilde{z})])$ is often known as the *certainty equivalence* of $\tilde{z}$ under the disutility u (Hershey and Schoemaker 1985, Zeldes 1989). For example, $\log\mathbb{E}[\exp(\cdot)]$ is a certainty equivalent operator.

The following representation theorem elucidates how we may construct a coherent bed shoratge metric using a convex expected disutility metric, a certainty equivalence or a convex risk measure (see for instance, Ben-Tal and Teboulle 2007, Föllmer and Schied 2002). We note that Brown and Sim (2009) have also provided a representation of a class of satisficing measure via a convex risk measure without the consideration of overloading. Our result is also broader in the sense that the properties of the functional μ in the representation theorem can be satisfied by a convex expected disutility metric, a certainty equivalence with convex disutility or a law invariant convex risk measure (Frittelli and Gianin 2005).

THEOREM 1 **(Representation)**. *The functional $\beta:\mathcal{Z}\to[0,1]$ is a* coherent bed shortage metric *if and only if there exists a riskiness evaluation functional $\rho:\mathcal{Z}\to[0,\infty]$ and an increasing calibrating function $\Phi:[0,\infty]\to[0,1]$ with $\Phi(0)=0$ and $\Phi(\infty)=1$ such that*

$$\beta(\tilde{z})=\Phi(\rho(\tilde{z}))$$

can be represented with

$$\rho(\tilde{z})=\inf\{\alpha>0\,|\,\mu_\alpha(\tilde{z})\le 0\} \tag{1}$$

for some family of functionals $\mu_\alpha(\tilde{z}):\mathcal{Z}\to\mathbb{R}$, $\alpha>0$, lower semi-continuous in $\tilde{z}$ and α, non-increasing in α, and satisfying the properties:
For all $\tilde{z},\tilde{z}_1,\tilde{z}_2\in\mathcal{Z}$,
(a) If $\mathbb{P}[\tilde{z}_1\ge\tilde{z}_2]=1$, then $\mu_\alpha(\tilde{z}_1)\ge\mu_\alpha(\tilde{z}_2)$ for all $\alpha>0$.
(b) If $\mathbb{E}[\tilde{z}]>0$, then $\mu_\alpha(\tilde{z})>0$ for all $\alpha>0$.
(c) $\mu_\alpha(0)=0$ for all $\alpha>0$.
(d) For all $\alpha>0$ and for all $\lambda\in[0,1]$, $\mu_\alpha(\lambda\tilde{z}_1+(1-\lambda)\tilde{z}_2)\le\max\{\mu_\alpha(\tilde{z}_1),\mu_\alpha(\tilde{z}_1)\}$.

Proof. The proof is presented in Appendix A. □

Theorem 1 reflects that the characterization of a coherent bed shortage metric is governed by the choice of the riskiness evaluation functional, $\rho(\cdot)$ as well as the calibrating function $\Phi(\cdot)$. The riskiness evaluation functional attempts to quantify how the risks of bed shortages would emerge and has an impact on tractability. The calibrating function maps the riskiness evaluation to $[0,1]$. With an appropriate choice, it would facilitate the interpretability of the metric in a sense similar to the BOR.

2.3. Aumann and Serrano (2008) riskiness index

If each functional $\mu_\alpha(\cdot)$ in Theorem 1 is the certainty equivalence of the exponential disutility with risk parameter α, then the corresponding riskiness evaluation functional $\rho(\cdot)$ becomes the well known Aumann and Serrano (2008) riskiness index, which we formally define as follows.

Definition 3 (Riskiness Index). The Aumann and Serrano (2008) riskiness index, $\rho : \mathcal{Z} \to [0, \infty]$, is defined as

$$\rho(\tilde{z}) = \inf \{\alpha > 0 \mid \mu_\alpha(\tilde{z}) \leq 0\},$$

where

$$\mu_\alpha(\tilde{z}) \triangleq \alpha \log \mathbb{E}[\exp(\tilde{z}/\alpha)]. \tag{2}$$

The riskiness index has an economic interpretation of being the reciprocal of the absolute risk aversion (ARA) of an individual with constant ARA who is indifferent to taking that gamble (Aumann and Serrano 2008).

It also relates to the other metrics previously mentioned, namely $\mathbb{P}[\tilde{z} > 0]$ and $\mathbb{E}[\tilde{z}^+]$. More specifically, the riskiness index yields bounds on these metrics. In the context of portfolio selection, Hall et al. (2015) provide the bound of $\mathbb{P}[\tilde{z} \geq \psi]$ as a function of the riskiness index for any level of violations, $\psi > 0$. Although the bounds are not necessarily tight, they provide assurance that the riskiness index is adequate in quantifying bed shortage risk.

Proposition 2. *For all $\psi \geq 0$, we have*

$$\mathbb{P}[\tilde{z} > \psi] \leq \exp\left(-\frac{\psi}{\rho(\tilde{z})}\right) \tag{3}$$

and

$$\mathbb{E}[(\tilde{z} - \psi)^+] \leq \frac{\rho(\tilde{z})}{e} \exp\left(-\frac{\psi}{\rho(\tilde{z})}\right). \tag{4}$$

Proof. The first inequality is an application of the Chernoff bound and is presented in Hall et al. (2015). As for the second, we note that $z^+ \leq \exp(z-1)$. By writing z as z/α and multiplying by α, we obtain, for all $\alpha > 0$, $z^+ \leq \alpha/e \cdot \exp(z/\alpha)$. Hence,

$$\mathbb{E}[(\tilde{z} - \psi)^+] \leq \frac{\rho(\tilde{z})}{e} \mathbb{E}\left[\exp\left(\frac{\tilde{z} - \psi}{\rho(\tilde{z})}\right)\right] \leq \frac{\rho(\tilde{z})}{e} \exp\left(-\frac{\psi}{\rho(\tilde{z})}\right).$$

□

Proposition 2 guarantees that the probability of ever larger shortages decreases exponentially. This decrease is controlled by the riskiness index. The lower the risk, the faster the rate of decay of the probability, *i.e.* the rate at which shortages will emerge. Likewise, $\mathbb{E}[(\tilde{z} - \psi)^+]$, the expected shortages beyond level $\psi > 0$, is also bounded.

Additionally, the conditional expected bed shortages $\mathbb{E}[(\tilde{z}-\psi)^+ \mid \tilde{z}>\psi] = \mathbb{E}[(\tilde{z}-\psi)^+]/\mathbb{P}[\tilde{z}>\psi]$ turns out to have a representation similar to the expression the riskiness evaluation functional in Theorem 1.

PROPOSITION 3.

$$\mathbb{E}[(\tilde{z}-\psi)^+ \mid \tilde{z}>\psi] = \inf\{\alpha>0 \;\mid\; \mathbb{E}[h((\tilde{z}-\psi)/\alpha)] \le \underbrace{\mathbb{E}[h(0)]}_{=1}\}$$

where

$$h(x) \triangleq \begin{cases} 1 & \text{if } x \le 0, \\ x & \text{otherwise.} \end{cases}$$

Proof. Observe that

$$\begin{aligned}
\mathbb{E}[(\tilde{z}-\psi)^+ \mid \tilde{z}>\psi] &= \mathbb{E}\left[(\tilde{z}-\psi)^+\right]/\mathbb{P}[\tilde{z}>\psi] && \text{: Note that } 0/0{=}0 \\
&= \inf\{\alpha>0 \;\mid\; \mathbb{E}\left[(\tilde{z}-\psi)^+\right]/\mathbb{P}[\tilde{z}>\psi] \le \alpha\} \\
&= \inf\{\alpha>0 \;\mid\; \mathbb{E}\left[(\tilde{z}-\psi)^+\right] \le \alpha\mathbb{P}[\tilde{z}>\psi]\} \\
&= \inf\{\alpha>0 \;\mid\; \mathbb{E}\left[((\tilde{z}-\psi)/\alpha)^+\right] \le \mathbb{P}[(\tilde{z}-\psi)/\alpha>0]\} \\
&= \inf\{\alpha>0 \;\mid\; \mathbb{E}\left[((\tilde{z}-\psi)/\alpha)^+\right] - \mathbb{P}[(\tilde{z}-\psi)/\alpha>0] + 1 \le 1\} \\
&= \inf\{\alpha>0 \;\mid\; \mathbb{E}\left[h((\tilde{z}-\psi)/\alpha)\right] \le 1\}.
\end{aligned}$$

□

The function $h(x)$ is locally continuous, convex and nondecreasing almost everywhere except at $x=0$. In particular, although the exponential disutility function, $\exp(x)$, is not the same as $h(x)$, both are nonnegative functions and their values coincide at $x=0$.

2.4. Bed shortage index

Although the Aumann and Serrano (2008) riskiness index is well-established in the economics community, it does not have a natural interpretation in the context of hospital bed shortages. For any appropriate choice of calibrating function, $\Phi(\cdot)$, the calibrated riskiness index, $\Phi(\rho(\tilde{z}))$ is a candidate for the coherent bed shortage metric. The goal of this section is to choose an appropriate calibrating function that yields a metric interpretable to healthcare practitioners and administrators. For this purpose, we propose the *bed shortage index*, or BSI for short, which is defined as follows:

DEFINITION 4 (BSI). The bed shortage index is the functional, $\beta : \mathcal{Z} \to [0,1]$,

$$\beta(\tilde{z}) = \Phi(\rho(\tilde{z})), \tag{5}$$

where the calibrating function is given by

$$\Phi(r) = \frac{1}{r\,(e^{1/r}-1)}$$

and $\rho(\tilde{z})$ is the Aumann and Serrano (2008) riskiness index.

Figure 1 illustrates how the riskiness index and the BSI are related in a one-to-one manner. While the riskiness index gives a *magnitude* of the risk involved, the BSI may be better understood in the sense of BORs. As we subsequently show in Theorem 2, it is calibrated to coincide with BOR when the daily arrivals in the hospital unit are Poisson distributed. Hence, in layman terms, the healthcare practitioner can interpret the BSI of a ward as the equivalent level of bed shortage risk experienced by a reference ward that is serving only Poisson arrivals, with its BOR being the same as the BSI.

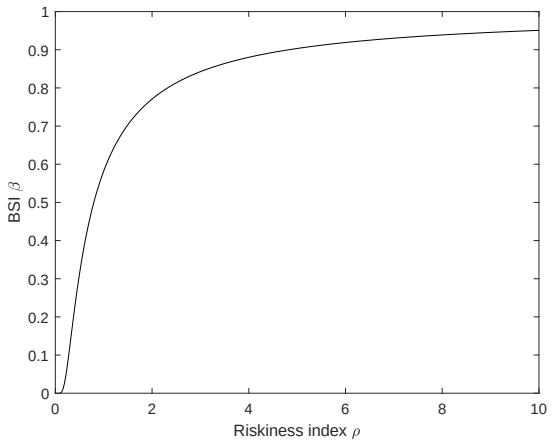


Figure 1 The calibrating function $\Phi(\cdot)$ mapping riskiness index ρ to BSI β

THEOREM 2 **(Interpretability)**. *Assume* $\mathbb{E}[\tilde{z}_t] \leq 0$. *If all the arrivals in a hospital unit follow a Poisson process with time-dependent rates* λ_t, *then the BSI and the BOR coincide on each day.*

Proof. By lower semi-continuity, we have

$$\begin{aligned}
\kappa_t = \mu_{\rho(\tilde{z}_t)}(\tilde{n}_t) &= \rho(\tilde{z}_t) \log \mathbb{E}[\exp(\tilde{n}_t/\rho(\tilde{z}_t))] \\
&= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \rho(\tilde{z}_t) \log \mathbb{E}[\exp(\tilde{y}_{\tau,s}/\rho(\tilde{z}_t))] \\
&= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \rho(\tilde{z}_t) \lambda_\tau q_s (e^{1/\rho(\tilde{z}_t)} - 1) \\
&= \rho(\tilde{z}_t) \mathbb{E}[\tilde{n}_t] \left(e^{1/\rho(\tilde{z}_t)} - 1\right)
\end{aligned}$$

Therefore, since $\kappa_t > 0$, $\beta(\tilde{z}_t) = \mathbb{E}[\tilde{n}_t/\kappa_t]$. □

The key result below illustrates that the BSI is in fact a risk-adjusted form of the BOR. It demarcates the conditions under which the BSI is larger, or smaller than BOR. We know that if all arrivals are Poisson, BSI equals to BOR which can be considered as the benchmark case. If

there is greater uncertainty, such as in the rates of arrivals, then the BSI would capture the risk and hence be larger than the BOR. Conversely, if all arrivals are certain and scheduled, then the BSI is smaller than the BOR, indicating less risk of bed shortages.

THEOREM 3 **(Risk awareness)**. *Under Assumption 1, if* $\mathbb{E}[\tilde{z}_t] \le 0$*, then BSI is a risk-adjusted form of the BOR. Suppose* $\tilde{n}_B(t)$ *is the number of elective patients on day* t*, and* $\tilde{n}_P(t)$ *is the number of emergency patients on day* t*. Then for* $\kappa_t > 0$*, the BSI satisfies:*

a. For the elective patients only, the BSI is smaller than BOR:

$$\beta(\tilde{n}_B(t) - \kappa_t) < \mathbb{E}[\tilde{n}_B(t)/\kappa_t]. \tag{6}$$

b. For the emergency patients only, the BSI is greater than BOR:

$$\beta(\tilde{n}_P(t) - \kappa_t) \ge \mathbb{E}[\tilde{n}_P(t)/\kappa_t]. \tag{7}$$

Proof. We first consider the case when there are only elective patients, i.e., $\tilde{z}_t = \tilde{n}_B(t) - \kappa_t$. According to Proposition 1, we know

$$\tilde{n}_B(t) = \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\ge 0}} \tilde{x}_{\tau,s},$$

where $\tilde{x}_{\tau,s} \sim \mathrm{Bin}(\eta_\tau, p_s)$. Let ρ be the riskiness index corresponding to BSI β, we have

$$\begin{aligned}
\rho(\tilde{z}_t)\log\mathbb{E}[\exp(\tilde{n}_B(t)/\rho(\tilde{z}_t))] &= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\ge 0}} \rho(\tilde{z}_t)\log\mathbb{E}[\exp(\tilde{x}_{\tau,s}/\rho(\tilde{z}_t))] \\
&= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\ge 0}} \rho(\tilde{z}_t)\log\left(1 - p_s + p_s e^{1/\rho(\tilde{z}_t)}\right)^{\eta_\tau} \\
&= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\ge 0}} \rho(\tilde{z})\eta_\tau \log\left(1 + \underbrace{p_s\left(e^{1/\rho(\tilde{z}_t)} - 1\right)}_{\ge 0}\right) \\
&< \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\ge 0}} \rho(\tilde{z}_t)\eta_\tau p_s\left(e^{1/\rho(\tilde{z}_t)} - 1\right) \qquad : \text{as } \log(1+x) \le x,\ \forall x \ge 0 \\
&= \rho(\tilde{z}_t)\mathbb{E}[\tilde{n}_B(t)]\left(e^{1/\rho(\tilde{z}_t)} - 1\right).
\end{aligned}$$

By lower semi-continuity of μ_α, we obtain that $\mu_{\rho(\tilde{z}_t)}(\tilde{z}_t) = 0$. Hence, $\kappa_t = \mu_{\rho(\tilde{z}_t)}(\tilde{n}) \le \rho(\tilde{z}_t)\mathbb{E}[\tilde{n}_B(t)]\left(e^{1/\rho(\tilde{z}_t)} - 1\right)$, or in other words, $\beta(\tilde{z}_t) \le \mathbb{E}[\tilde{n}_B(t)/\kappa_t]$.

We then consider the case when there are only emergency arrivals, i.e., $\tilde{z}_t = \tilde{n}_P(t) - \kappa_t$. According to Proposition 1, we know

$$\tilde{n}_P(t) = \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\ge 0}} \tilde{y}_{\tau,s},$$

where $\tilde{y}_{\tau,s} \sim \text{Pois}(\tilde{\lambda}_\tau q_s)$. Then, we have

$$\begin{aligned}
\rho(\tilde{z}_t)\log\mathbb{E}[\exp(\tilde{n}_P(t)/\rho(\tilde{z}_t))] &= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \rho(\tilde{z}_t)\log\mathbb{E}[\exp(\tilde{y}_{\tau,s}/\rho(\tilde{z}_t))] \\
&= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \rho(\tilde{z}_t)\log\mathbb{E}[\exp(\tilde{\lambda}_\tau q_s(e^{1/\rho(\tilde{z}_t)}-1))] \\
&\geq \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \rho(\tilde{z}_t)\mathbb{E}[\tilde{\lambda}_\tau q_s]\left(e^{1/\rho(\tilde{z}_t)}-1\right) \qquad \text{: Jensen's inequality} \\
&= \rho(\tilde{z}_t)\mathbb{E}[\tilde{n}_P(t)]\left(e^{1/\rho(\tilde{z}_t)}-1\right)
\end{aligned}$$

Therefore, by lower semi-continuity, $\kappa_t = \mu_{\rho(\tilde{z}_t)}(\tilde{n}) \geq \rho(\tilde{z}_t)\mathbb{E}[\tilde{n}_P(t)]\left(e^{1/\rho(\tilde{z}_t)}-1\right)$ or in other words, since $\kappa_t > 0$, $\beta(\tilde{z}_t) \geq \mathbb{E}[\tilde{n}_P(t)/\kappa_t]$. □

Table 1 compares the BSI to other metrics. As far as we know, BSI is the only metric that fulfils all the desired properties of a coherent bed shortage metric. BOR, despite being interpretable and tractable, violates the risk-free property and hence does not capture the various level of risks associated with bed shortages well. Although the probability of bed shortages and expected shortages are interpretable, they are computationally intractable and do not identify situations of overloading. The calibrated riskiness index has most of the salient properties, but without the appropriate calibrating function, it may not be easily understood by healthcare practitioners.

Table 1 Various bed shortage metrics and their properties

	BOR	Prob. of shortages	Expected shortages	Calibrated Riskiness Index	BSI
Monotonicity	✓	✓	✓	✓	✓
Risk-free	✗	✓	✓	✓	✓
Overloading	✓	✗	✗	✓	✓
Quasiconvexity	✓	✗	✓	✓	✓
Tractability	✓	✗	✗	✓	✓
Interpretability	✓	✓	✓	✗	✓

It is worth mentioning that the BSI is not the only class of proper bed shortage metric. The differentiating consideration, however, is its tractability and interpretability. In particular, the bed shortage position $\tilde{z}$ is a composite random variable comprising a sum of independently distributed, but not necessarily identical, random variables whose distributions endogenously depend on the decision variables to be optimized. Moreover, it is generally intractable to evaluate the distribution of such a random variable to arbitrary precision (see, for instance, Khachiyan 1989, Nemirovski

and Shapiro 2006). Hence, although we can replace $\mu_\alpha(\tilde{z})$ by the popular conditional value-at-risk (CVaR) to yield a coherent bed shortage metric by Theorem 1, we do not know how we could formulate this as a tractable optimization problem (Rockafellar et al. 2000). In contrast, the BSI incorporates higher-order statistical information while providing a closed-form expression for $\mu_\alpha(\tilde{z})$ in terms of its parameters. We see this later in Theorem 4. Henceforth, we adopt this metric for evaluating and minimizing the risk associated with bed shortages.

3. Descriptive analytics

In this section, we show how the BSI for a ward may be computed and we subsequently compare it against the BOR and other traditional metrics. Theorem 4 provides an exact closed-form solution for the BSI. To facilitate understanding, we focus on a single ward with elective and emergency patients in the steady state. However, the results can be easily extended to a ward with multiple types of patients (see Appendix C.1 for more discussions on patient segmentation), and even in the transient state if Assumption 2 is relaxed.

3.1. Steady-state formulation

To understand how capacity can be planned, we wish to study bed shortages from a steady-state perspective. To this end, we define what it means to be operating in the steady-state context.

Assumption 2 **(Steady-state)**. *Given a time window T, we say that in the steady-state, the following variables are periodic, i.e., for, $t, t' \in \bar{\mathcal{T}}$ such that $t \equiv t' (mod\ T)$, we have,*

1. *Periodic quota: $\eta_t = \eta_{t'}$;*
2. *Periodic arrival rates: $\lambda_t = \lambda_{t'}$; and,*
3. *Periodic capacity: $\kappa_t = \kappa_{t'}$.*

As such, it suffices to consider these variables, only over the planning window $\mathcal{T}$.

In the presence of elective admissions, it is difficult to efficiently compute expected shortages $\mathbb{E}[\tilde{z}_t^+]$ and the probability of shortages $\mathbb{P}[\tilde{z}_t > 0]$ at steady-state. We have to resort to simulations to do so, as in the numerical illustration later. Instead, the BSI can be efficiently computed. This is our key result – μ has a closed-form expression:

Theorem 4 **(Tractability)**. *Under Assumptions 1 and 2, for $t \in \mathcal{T}$, μ_{α_t} can be evaluated in closed-form as*

$$\mu_{\alpha_t}(\tilde{z}_t) = \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} A_{\tau+jT}(\alpha_t)\eta_{t-\tau} + \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} \alpha_t \Lambda_{t-\tau}(q_{\tau+jT}(e^{1/\alpha_t} - 1)) - \kappa_t, \tag{8}$$

where $A_s(\alpha) \triangleq \alpha \log(1 - p_s + p_s e^{1/\alpha})$. Since $\mu_{\alpha_t}(\tilde{z}_t)$ is non-increasing in α_t, the corresponding BSI, $\beta(\tilde{z}_t)$ can be tractably evaluated by performing binary search on α_t until $\alpha_t = \alpha_t^*$ for which $\mu_{\alpha_t^*}(\tilde{z}_t) = 0$ and $\beta(\tilde{z}_t) = \Phi(\alpha_t^*)$.

Proof. For fixed α_t, since μ_{α_t} is additive over independently distributed random variables,

$$\mu_{\alpha_t}(\tilde{z}_t) = \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \mu_{\alpha_t}(\tilde{x}_{\tau,s}) + \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \mu_{\alpha_t}(\tilde{y}_{\tau,s}) - \kappa_t.$$

The first sum may be evaluated as follows.

$$\begin{aligned}
\sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \alpha_t \log\Big(\mathbb{E}[\exp(\tilde{x}_{\tau,s}/\alpha_t)]\Big) &= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \alpha_t \eta_\tau \log\left(1 - p_s + p_s e^{1/\alpha_t}\right) && (9)\\
&= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} A_s(\alpha_t)\eta_\tau \\
&= \sum_{\tau\geq 0} \eta_{t-\tau} A_\tau(\alpha_t) \\
&= \sum_{\tau=0}^{T-1} \eta_{t-\tau} \sum_{j=0}^{\infty} A_{\tau+jT}(\alpha_t) && (10)\\
&= \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} A_{\tau+jT}(\alpha_t)\eta_{t-\tau}
\end{aligned}$$

We used Proposition 1(a) for (9) in evaluating its moment generating function, and for (10), we grouped terms using periodicity. This recovers the x-term in (8). We can repeat the procedure for the y-term using the moment generating function for $\tilde{y}_{t,s} \sim \text{Pois}(\tilde{\lambda}_t q_s)$ in (11):

$$\begin{aligned}
\sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \alpha_t \log\Big(\mathbb{E}[\exp(\tilde{y}_{\tau,s}/\alpha_t)]\Big) &= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \alpha_t \log\Big(\mathbb{E}[\exp(\tilde{\lambda}_\tau q_s(e^{1/\alpha_t}-1))]\Big) && (11)\\
&= \sum_{\substack{(\tau,s)\,:\,\tau+s=t \\ s\geq 0}} \alpha_t \Lambda_\tau(q_s(e^{1/\alpha_t}-1)) \\
&= \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} \alpha_t \Lambda_{t-\tau}(q_{\tau+jT}(e^{1/\alpha_t}-1)).
\end{aligned}$$

□

The resultant form can be tractably computed, because:

1. Given fixed α_t, the equation (8) is affine in η_t and κ_t.
2. The equation (8) is non-increasing in α_t for fixed η_t.

This is important. The first observation allows us to optimize the BSI using only linear constraints. Moreover, the second observation implies that the minimum value of α_t for which $\mu_{\alpha_t}(\tilde{z}_t) \leq 0$ occurs when it is tight. We see its exact ramifications later.

In the following corollary, we provide explicit solutions for three special cases based on Theorem 4. The first case is with known arrival rate, i.e., $\tilde{\lambda}_t = \lambda_t$. The second case is a two-point distribution, $\tilde{\lambda}_t = \lambda_t + \tilde{d}_t\Delta$, where $\tilde{d}_t$ is i.i.d. Bernoulli with probability $\bar{p}$. This represents the situation where

there is the possibility of demand surges of rate Δ on top of the ordinary arrival rate λ_t. The surges occur with probability $\bar{p}$. For the third case, we assume that $\tilde{\lambda}_t$ follows a exponential distribution with mean λ_t. Of course, explicit solutions for other distributions could be derived in a similar way.

COROLLARY 1. *Under Assumptions 1 and 2, for $t \in \mathcal{T}$, μ_{α_t} can be re-formulated as follows:*

a. If $\tilde{\lambda}_t = \lambda_t$, we have

$$\mu_{\alpha_t}(\tilde{z}_t) = \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} A_{\tau+jT}(\alpha_t)\eta_{t-\tau} + \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} \alpha_t \left(e^{1/\alpha_t} - 1\right) \lambda_{t-\tau} q_{\tau+jT} - \kappa_t. \tag{12}$$

b. If $\tilde{\lambda}_t = \lambda_t + \tilde{d}_t \Delta$, $\tilde{d}_t \sim B(\bar{p})$, we have

$$\begin{aligned} \mu_{\alpha_t}(\tilde{z}_t) = &\sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} A_{\tau+jT}(\alpha_t)\eta_{t-\tau} + \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} \alpha_t \left(e^{1/\alpha_t} - 1\right) \lambda_{t-\tau} q_{\tau+jT} \\ &+ \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} \alpha_t \log\left(1 - \bar{p} + \bar{p} \exp((e^{1/\alpha_t} - 1) q_{\tau+jT} \Delta)\right) - \kappa_t. \end{aligned} \tag{13}$$

c. If $\tilde{\lambda}_t \sim Exp(\lambda_t)$, we have

$$\mu_{\alpha_t}(\tilde{z}_t) = \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} A_{\tau+jT}(\alpha_t)\eta_{t-\tau} + \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} -\alpha_t \log\left(1 - (e^{1/\alpha_t} - 1)\lambda_{t-\tau} q_{\tau+jT}\right) - \kappa_t. \tag{14}$$

3.2. A numerical example: descriptive metrics in a ward

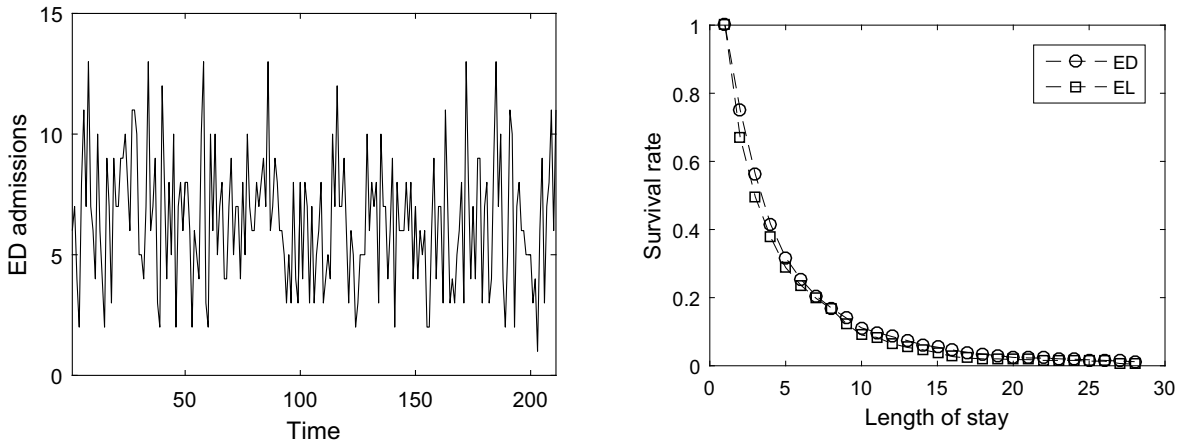
We illustrate the application of BSI using real data drawn from a ward in a general hospital. We first describe the data and the studied case. Using the closed-form expression (8) in Theorem 4, we are able to calculate the BSI under different arrival regimes. We present three cases where the BSI is larger, equal, or smaller than BOR, which illustrates that the BSI is indeed the risk-adjusted form of BOR. We then check the validity of Assumptions 1 and 2, and estimate the arrival rates of emergency patients using the data. In order to give a more intuitive understanding of the BSI, we simulate classical performance measures such as probability of shortages and expected shortages. We compare the simulated results with the BSI, and find that they are consistent as risk measures. Finally, we apply the BSI to study the impact of long stayers.

Data description. There are 44 beds in the ward, and we studied them over 210 days. The data captures the admission and discharge dates of each patient. Patients were admitted from both the ED and from elective surgeries, with average LOS around 4.73 days. Figure 2(a) shows the daily emergency admissions over the studied period, and its weekly pattern supports our choice of $T = 7$. Figure 4(a) plots the average daily emergency admission pattern within a week. Figure 2(b) plots the survival rates of emergency and elective patients, *i.e.*, the proportion of patients who remain warded after s days, p_s and q_s. While the majority of patients were discharged by the

fifth day, a small but significant proportion remain warded for more than 2 weeks. Also, we notice a difference in survival rates between patients admitted through the emergency and elective routes – on average, the LOS of emergency patients of about 4.82 days is longer than that of elective patients at 4.24 days. Here and later, we assume that all scheduled elective patients turn up and stay for at least a day, *i.e.*, $p_0 = 1$. As emergency patients are normally admitted within the day of arrival, we use the daily admissions to approximate the daily arrivals in the following analysis.

Figure 2 ED admissions and survival rates of emergency and elective patients

(a) ED Admissions (b) Survival rates



BSI calculation. Under Assumptions 1 and 2, and given the arrival process and the survival probabilities of LOS, we can compute the BOR, $\mathbb{E}[\tilde{n}_t/\kappa]$, and the BSI using (5) and (8). To illustrate how the BSI captures the risk, we conduct a numerical analysis of three wards (Ward S, Ward P and Ward D) under different arrival regimes (Poisson with random or constant rate, versus Deterministic). Mean daily arrivals are assumed at (10, 10, 8, 8, 9, 8, 6). While it was designed for these Wards to have the same BOR, their BSIs differ as Figure 3 illustrates. When the arrivals are Poisson with constant rates, by Theorem 3, values for the BSI and the BOR should coincide. Under this traditional assumption, BOR may be able to capture the risk as well as BSI. However, when there is uncertainty in the rate of arrival, it will fail to do so. Indeed, Ward S experiences greater risks of bed shortages than Ward P, while Ward P experiences greater risks than Ward D. This underscores how our proposed BSI can extend the descriptive power of the popular BOR as the proxy for quantifying bed shortage risks.

In practice, hospital managers usually use BOR to evaluate the risk of bed shortages. Unfortunately, the BOR is less sensitive to uncertainty. For instance, one would expect to see higher risks of shortages in a ward admitting only emergency patients, than one where arrivals are only from

elective patients and completely deterministic. Yet, they will surely have the same BOR. The BSI captures this difference – wards with the same BOR can have significantly different BSIs. Often, the BSI is associated with the number of admissions in that day – the higher the admissions, the higher the metric. However, this is not always the case due to the confounding effect of the LOS of patients. As Figure 3 illustrates, the second highest BSI falls on Wednesdays. This is because the arrivals on Sundays are lower and so, despite the high arrivals on Mondays, the wards are also comparatively emptier. This relieves the pressure on shortages on Mondays.

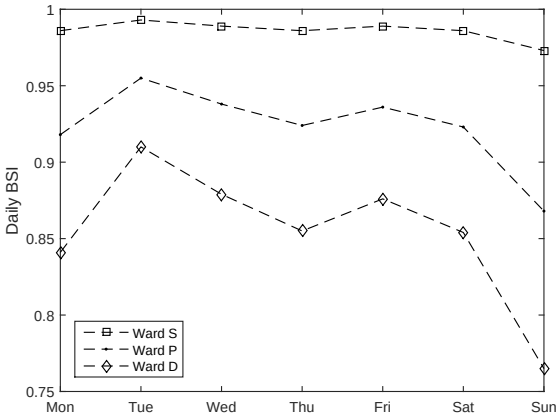


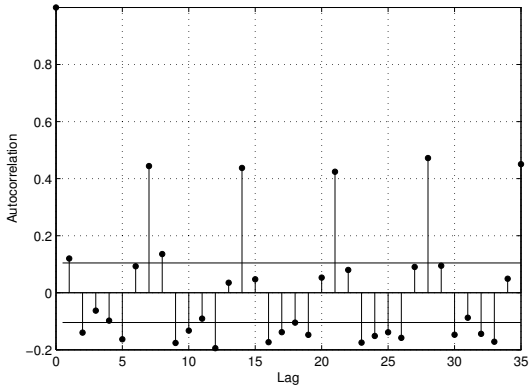
Figure 3 Daily BSI under different arrival processes

Simulated performance. Here, we first fit distributions to our arrival data, then calculate the BSI under the fitted distributions, and simulate the performance accordingly. The quantile-quantile plot in Figure 4(b) supports the assumption to model daily emergency arrivals via a Poisson distribution with constant rate. The null hypothesis that daily emergency admissions follows a Poisson distribution is moreover not rejected at the 5% significance level using a χ^2 test. Henceforth, we adopt the Poisson assumption with constant rates in this analysis unless otherwise specified. These rates are estimated at (7.66, 7.89, 6.49, 6.25, 7.09, 5.68, 4.11) from Monday through Sunday. We notice that the arrival rates are higher on weekdays, as is also observed by local physicians.

Given these estimated daily arrival rates, we simulate the Poisson arrival process for a length of 1 million weeks, *i.e.* 7 million days. For each patient, we also simulate the LOS according to the empirical distribution. The simulation warm-up period is the first 100,000 weeks, which is not included in the results. Table 2 shows the simulated probability of shortages and expected shortages, illustrating the consistency of the BSI with classical performance measures.

Figure 4 The daily number of ED admissions in 210 days

(a) Autocorrelation of emergency admissions



(b) Q-Q plot

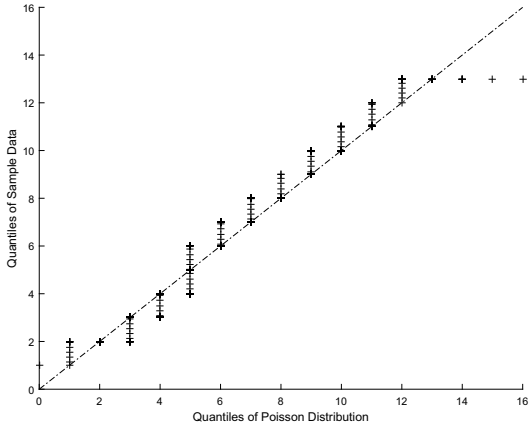


Table 2 Summary statistics of daily bed shortages based on emergency arrivals (Poisson arrival)

		Analytical results			Simulated results		
	λ_t	BOR	BSI	$\rho(\tilde{z}_t)$	$\mathbb{P}[\tilde{z}_t > 0]$	$\mathbb{E}[\tilde{z}_t^+]$	$\mathbb{E}[\tilde{z}_t^+ \vert \tilde{z}_t > 0]$
Mon	7.660	0.664	0.664	1.300	0.004	0.010	2.475
Tue	7.887	0.700	0.700	1.484	0.010	0.026	2.687
Wed	6.491	0.695	0.695	1.457	0.009	0.023	2.679
Thu	6.245	0.686	0.686	1.404	0.007	0.018	2.608
Fri	7.093	0.697	0.697	1.467	0.009	0.025	2.684
Sat	5.679	0.675	0.675	1.351	0.005	0.014	2.534
Sun	4.113	0.623	0.623	1.133	0.001	0.003	2.270

Impact of long stayers. Health systems across the world have implemented the concept of step-down care and transitional care. The motivation is that acute or general hospitals may not be the best setting for patients requiring a lower level of, but longer term care. Such patients include those suffering from chronic illnesses, or whom may be taking longer to recover from severe incidents. More often than not, these patients are advanced in age. Patients requiring less intensive care may divert treatment resources from more pressing cases that acute hospitals receive. Instead, nursing homes, physiotherapists, and other community care-providers might be better placed to provide for these needs, both in terms of quality and frequency of care. In Singapore, community hospitals, which cater to the provision of such care, have sprung up of late. It is anticipated that these community hospitals will stand to gain from economies of scale, and operate on a lower budget than acute hospitals. In particular, we seek to quantify the effect of a policy which transfers

'long-stayers' (here, referred to as patients with LOS exceeding a certain threshold L) to community hospital and other facilities.

Figure 5 Daily BSI under different maximal LOS

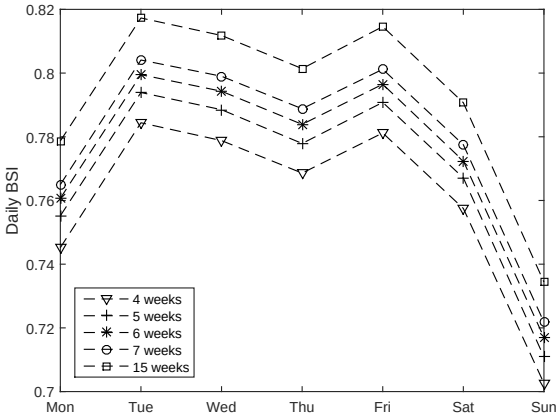
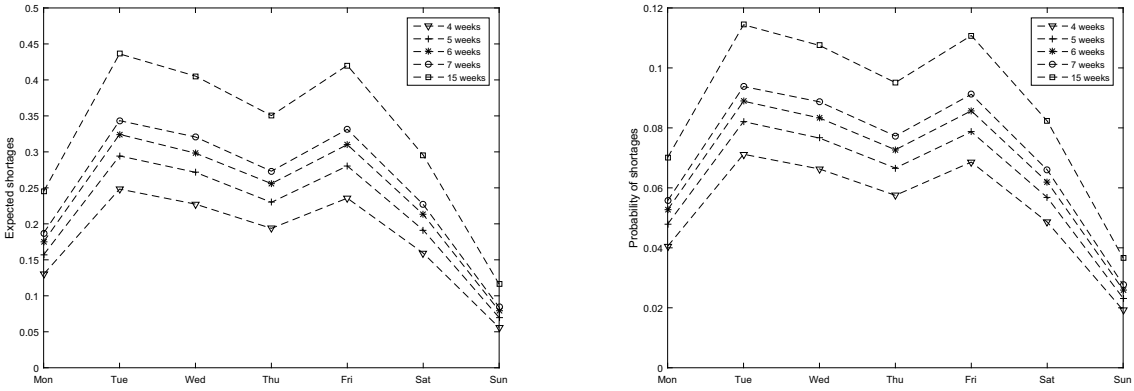


Figure 6 Daily expected shortages and probability of shortages under different maximal LOS

(a) Expected shortages (b) Probability of shortages



In this analysis, L is set at 4 weeks. Long-stayers comprise only about 1% of our samples. Their longest LOS is observed at 105 days (15 weeks). To examine their impact, we progressively reduce the truncation L, by allowing the maximal LOS to vary from 4 weeks upwards till it reaches 15 weeks. This would represent the effect of a transfer policy to move out patients after L days.

Let us assume that there is just one elective admission per day, *i.e.*, 7 within the week. Figure 5 illustrates what happens when the ward continues to retain patients beyond the 4-week mark. While accounting for only 1% of the patients, the presence of long-stayers increases the BSI by around

4% daily. This appears to support the hypothesis by practitioners that long-stayers disproportionately affect the level of bed shortages. Figure 6 shows the simulated daily expected shortages and probabilty of shortages within the week under different maximal LOS. These results are consistent with Figure 5. We return to this analysis later after introducing a model for optimizing elective admissions in the steady state. At that point, we examine their impact *under optimal policy*.

4. Prescriptive analytics

Prescriptive analytics entails optimization techniques to efficiently and effectively determine the best performing solution over a criterion among possibly astronomically large number of alternatives. Hence, it is important for such a criterion to be amenable to computation. As mentioned in our earlier illustration, the probability of bed shortages and the expected shortages are not easily computable up to arbitrary accuracy, let alone employed as the criterion to be optimized. Instead, we employ the BSI to examine how the risk of bed shortages may be minimized.

In this spirit, let $\mathcal{I}$ represent some collection of wards. Depending on the problem setting, $\mathcal{I}$ can represent different objects. If the administrator is looking at optimizing the scheduling within a group of wards with different classes of patients, then $\mathcal{I}$ can represent wards grouped by each class of patients. If the administrator is studying the allocation of bed capacity across different departments, then $\mathcal{I}$ represents the different departments in the hospital. At the broadest level, if a health planner is looking at the capacity of hospitals within a region, then $\mathcal{I}$ can represent different hospitals. Each element in $\mathcal{I}$ (class of patients, ward, department or hospital) is assumed to be managed separately, with an overall administrator that might decide on some operating parameters for them.

In the most general sense, this administrator can have three decision variables – elective admission quota $\boldsymbol{\eta} = (\eta_t^i)_{t\in\mathcal{T},i\in\mathcal{I}}$, distribution of bed capacity $\boldsymbol{\kappa} = (\kappa_t^i)_{t\in\mathcal{T},i\in\mathcal{I}}$, and the rate of emergency arrivals $\boldsymbol{\lambda} = (\lambda_t^i)_{t\in\mathcal{T},i\in\mathcal{I}}$. At this point, it may seem counter-intuitive for the arrival rate to be a decision variable, but we see in Section 5.4 later, how this could be applied. In totality, let $\boldsymbol{w} = (\boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\lambda})$ represent all the decision variables. Denote by $\mathcal{W}$ be the feasible set for $\boldsymbol{w}$. Usually, a hospital administrator would consider only one or two sets of decision variables, which could be done by setting other decision variables as known parameters.

At the ward or hospital level, one may assume that arrival rates $\boldsymbol{\lambda}$ are fixed and consider the following set of constraints:

- *Total capacity constraints*: $\sum_{i\in\mathcal{I}} \kappa_t^i \le C,\ \forall t \in \mathcal{T}$. Total bed capacity must be bounded by the total beds in the entire hospital, assumed to be constant over the planning horizon.
- *Resourcing constraints*: $\sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}} m_t^i \kappa_t^i \le R$. These constraints involve the other resources required in supporting the operations of the wards, such as the staffing of physicians and care-providers, or the financial costs of medical equipment, etc...

- *Service continuity constraints*: Each ward i needs to achieve a minimal target of service continuity, in terms of the usual scheduled rate of surgeries. Specifically, $\sum_{t\in\mathcal{T}} \eta_t^i \geq S^i$, $\forall i \in \mathcal{I}$.

High elective admissions are generally more profitable for a hospital, but it would impact the capacity of the hospital for emergency admissions. Moreover, physicians are often attracted to a hospital with high elective admissions, so that they have more opportunities to hone their surgical skills. A private hospital would generally desire to maximize the number of elective admission to maximize profit as well as attract good physicians as long as the risk of bed shortages is acceptable. As we have explained, the healthcare practitioner can interpret the BSI as the risk of bed shortages associated with a ward with the same BOR serving only Poisson distributed emergency arrivals. When maximizing the number of elective admissions, this interpretation is useful for healthcare administrators in specifying the desired level of bed shortage risk by benchmarking against such a ward. On the other hand, a public hospital faces constraints on elective admission and the admission quota is determined by the government (see Meng et al. 2015). Hence, we present different optimization models, which can be used in different contexts.

4.1. Maximizing throughput

Our first model attempts to maximize throughput, *i.e.* total elective admissions, subject to the BSI remaining bounded below a certain threshold. In this case, $\mathcal{I}$ represents a collection of wards. This turns out to be a linear model.

At all times, the hospital should have a control over the maximum risk of bed shortage that is permissible – the planner sets a maximum acceptable BSI $\epsilon_t^i \in [0,1]$ over time period t and for different wards i. The dependence on wards i can be explained as follows: Suppose some particular i represents the Intensive Care Unit say, then it can be imagined that its allowed BSI ϵ_t^i would differ from that of a normal ward where the consequences of bed shortages are less perilous. This leads to the model:

$$\begin{aligned} \max \ & \sum_{i\in\mathcal{I}} \delta^i \sum_{t\in\mathcal{T}} \eta_t^i \\ \text{s.t. } & \beta(\tilde{z}_t^i(\boldsymbol{w})) \leq \epsilon_t^i, \ \ \forall t \in \mathcal{T}, \forall i \in \mathcal{I} \\ & \boldsymbol{w} \in \mathcal{W}. \end{aligned} \tag{15}$$

Here, δ^i can be a flexible weight (such as expected profit per elective) for each ward i. In the simplest case, $\delta^i = 1$. Theorem 4 grants us the reformulation:

Corollary 2. *Under Assumptions 1 and 2, the throughput maximization model (15) has the following linear reformulation:*

$$\begin{aligned} \max \ & \sum_{i\in\mathcal{I}} \delta^i \sum_{t\in\mathcal{T}} \eta_t^i \\ s.t. \ & \textstyle\sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} A_{\tau+jT}^i(\alpha_t^i)\eta_{t-\tau}^i + \sum_{\tau=0}^{T-1} \sum_{j=0}^{\lfloor L/T \rfloor} \alpha_t^i \Lambda_{t-\tau}^i (q_{\tau+jT}^i (e^{1/\alpha_t^i} - 1)) \leq \kappa_t^i, \ \ \forall t \in \mathcal{T}, \forall i \in \mathcal{I} \\ & (\boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\lambda}) \in \mathcal{W} \end{aligned} \tag{16}$$

where $\alpha_t^i \in [0,\infty]$ *corresponds to* $\epsilon_t^i = 1/\alpha_t^i(e^{1/\alpha_t^i}-1)$.

4.2. Minimizing overall risk

An alternative would be to minimize the BSIs, $(\beta(\tilde{z}_t^i(\boldsymbol{w})))_{t\in\mathcal{T},i\in\mathcal{I}}$, subject to the operating context, *e.g.* a minimal number of elective patients that must be scheduled. The core challenge is that this set of BSIs is not a single objective function to be minimized. In such a case, one often attempts optimizing a composite measure. Two approaches are common in the literature. The first minimizes the sum of BSIs across wards and time periods. However, this may lead to an intractable nonlinear and non-convex formulation which we do not know how to optimize efficiently. The significance of summing the BSIs may also be difficult to interpret. Additionally, the risk may not be distributed fairly across wards or across time periods – the optimal solution may find it justified to increase severely the risk of a single ward in exchange for lower risk levels in other wards. To address this, the second approach is to minimize the worst-case BSI over all time periods and wards. Such min-max fairness criterion would typically lead to a more balanced allocation of risks across wards and time periods.

Conceivably, even if the worst-case BSI were minimized at a particular ward and time period, it may well be possible to further reduce the risk in other wards and time periods without compromising the min-max fairness criterion. As an extension, we adopt a lexicographic minimization (or 'lex min' for short). We detail in Appendix B its modelling and algorithmic approaches. This paradigm leads to the general model:

$$\begin{aligned} &\text{lex min } (\beta(\tilde{z}_t^i(\boldsymbol{w})))_{t\in\mathcal{T},i\in\mathcal{I}} \\ &\text{s.t. } \boldsymbol{w}\in\mathcal{W}. \end{aligned} \tag{17}$$

Again, Theorem 4 grants us the privilege of the following reformulation:

Corollary 3. *Under Assumptions 1 and 2, the risk minimization model (17) has the following equivalent reformulation:*

$$\begin{aligned} &\text{lex min } \boldsymbol{\alpha} \\ &\text{s.t. } \textstyle\sum_{\tau=0}^{T-1}\sum_{j=0}^{\lfloor L/T\rfloor} A_{\tau+jT}^i(\alpha_t^i)\eta_{t-\tau}^i + \sum_{\tau=0}^{T-1}\sum_{j=0}^{\lfloor L/T\rfloor}\alpha_t^i\Lambda_{t-\tau}^i(q_{\tau+jT}^i(e^{1/\alpha_t^i}-1)) \le \kappa_t^i,\ \forall t\in\mathcal{T},\forall i\in\mathcal{I} \\ &\quad (\boldsymbol{\eta},\boldsymbol{\kappa},\boldsymbol{\lambda})\in\mathcal{W},\boldsymbol{\alpha}>0. \end{aligned} \tag{18}$$

This model is flexible. Capacity κ_t^i may depend on t. This means that the hospital may adaptively move beds across wards over the course of a week. Alternatively, the model can be modified to include a ward with 'back-up capacity' which houses patients overflowing from other wards.

4.3. Numerical examples

In this section, we illustrate how to apply the throughput maximization model (16) as well as the risk minimization model (18) through numerical examples when emergency arrivals follow Poisson distributions with known rates. We will show an example in Section 5.3 when emergency arrivals follow doubly stochastic Poisson. By Corollary 2, the former reduces to a linear optimization problem, while the latter is implemented via Algorithm 1 in Appendix B. Using simulations as described earlier, we study their performance, such as the probabilities of different shortage levels and expected shortages. We also use our models to examine the impact of long stayers numerically.

Maximizing ward throughput. In the case of maximizing the number of elective admissions, we dictate the level ϵ which the BSI has to be controlled under, i.e., $\beta(\tilde{z}_t) \leq \epsilon$ for all days $t \in \mathcal{T}$. ϵ is then varied. The optimal scheduling solution is shown in Table 3. When the maximum BSI level ϵ is allowed to increase from 0.75 to 0.95, the maximum allowed number of elective admissions S correspondingly increases from 2 to 17 per week. Also, admissions are preferentially scheduled on weekends, when emergency arrivals are lower. In this sense, the optimal policy *smooths* admissions to equalize the risks across days, in anticipation of shortages over the week.

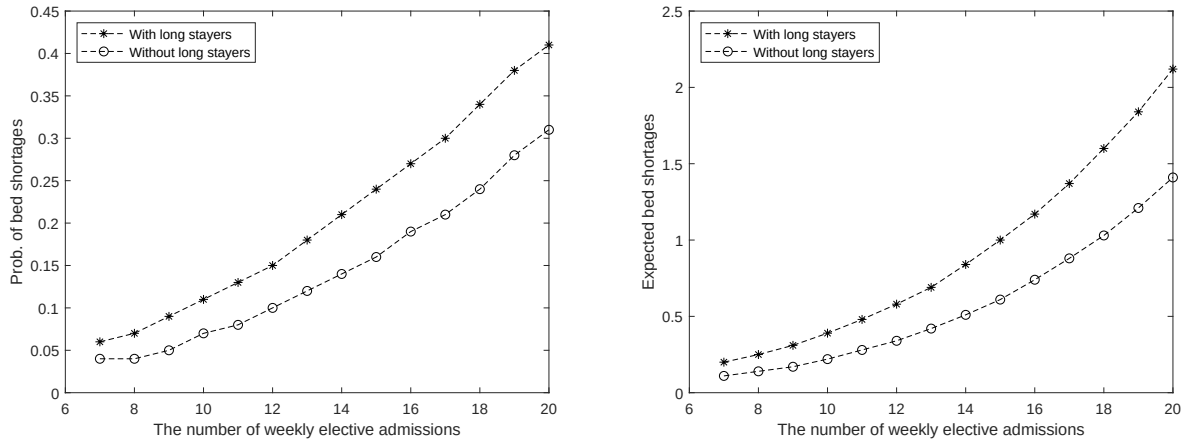
In Table 3, we also verify the performance of the optimal policy using simulations. The table illustrates that when the maximum BSI level was allowed to increase from 0.75 to 0.95, the average BOR rose from 0.73 to 0.94. The probability of bed shortages rose from 0.02 to 0.30 and the expected number of bed shortages surged from 0.06 to 1.37.

Recall that the BSI is interpreted as the equivalent level of bed shortage risk experienced by a reference ward with the same BOR serving only Poisson arrivals. This would help healthcare practitioners and administrators in specifying the desired threshold, ϵ. For instance, it is a common wisdom is to keep the BOR under 85% (Green 2002). Given an optimal elective scheduling solution, its performance can be evaluated on other traditional metrics such as the probability of bed shortages via simulations, such as in Table 3. From this, the ward manager may decide upon the configuration which best fits his purposes.

Minimizing bed shortage risk. In this model, we instead fix $S = \sum_{t \in \mathcal{T}} \eta_t$, the number of elective patients that must be admitted per week. The model then seeks an optimal policy scheduling these S elective patients, so that the BSIs over the week are minimized. When elective throughput S is increased from 7 to 20, the optimal worst-case BSI rises from 0.80 to 0.99. Similar to the previous model, more patients are admitted over the weekends. Figure 7 shows the risk of shortages based on our simulations. From our analysis, the trade-off for maximizing BOR can be harsh. Increasing total elective throughput S from 7 to 20 raises the probability of bed shortages

Table 3 The optimal elective scheduling and service level given maximum BSI level ϵ

	Optimal elective scheduling								Ave.	Prob. of	Expected Shortages	
ϵ	η_1	η_2	η_3	η_4	η_5	η_6	η_7	S	BOR	Shortages	Unconditional	Conditional
0.75	0	0	0	1	0	1	0	2	0.73	0.02	0.06	2.93
0.80	0	0	0	1	0	3	2	6	0.79	0.05	0.16	3.20
0.85	0	0	1	2	1	0	6	10	0.85	0.11	0.40	3.62
0.90	1	0	2	2	1	3	5	14	0.90	0.21	0.84	4.09
0.95	1	1	2	3	2	3	5	17	0.94	0.30	1.37	4.55



(a) Prob. of Shortages (b) Expected Shortages

Figure 7 The simulated probability of bed shortages and expected bed shortages under BSI minimization model given the weekly elective quota with or without long stayers

from 0.06 to 0.41 and expected shortages from 0.20 to 2.12. See Tables 7 and 8 in the appendix for more details.

This very short analysis identifies a key piece of understanding: shortages are unavoidable, even under optimal control. Instead of trying to increase bed capacity, it may be more effective to devise strategies to pool together the risks across different wards to alleviate these shortages. One such approach may be to create an 'overflow ward' which houses patients who could not be assigned a bed from any department.

Examining the impact of long-stayers. We had studied earlier how the bed shortage risk changes with the exclusion of long-stayers. Here, we examine their effect on the optimal elective throughput. To do so, we compare between two policies: one with long-stayers, and another where the long-stayers are transferred to community hospitals after 28 days (4 weeks).

Using our proposed algorithm, we obtain the optimal scheduling policy without long-stayers and its daily BSI $\beta(\tilde{z}_t)$, for different levels of elective throughput S. Similarly, we plot the simulated performance in Figure 7, which seems to out-performs the case with long stayers. We observe that this is due to the lower daily BSI $\beta(\tilde{z}_t)$ under a policy where long-stayers are transferred. See Tables 9 and 10 in the appendix for more details.

We also take an alternative approach to study this problem: How much capacity may be freed up for new patients if we transferred all long-stayers to community hospitals at the 4-week mark? To do so, we utilize the throughput maximizing model (16), by varying the desired risk level or BSI threshold ϵ. We plot in Table 4, the maximum elective throughput S if there were no long-stayers, for each ϵ. By comparing Tables 3 and 4, the capacity freed up by this transfer policy works out to around 2 to 3 new patients, while preserving the same level of performance. This amounts to at least a 15% increase in elective quota.

Table 4 The optimal elective scheduling and service level without long-stayers given BSI threshold ϵ

	Optimal elective scheduling								Average	Prob. of	Expected Shortages	
ϵ	η_1	η_2	η_3	η_4	η_5	η_6	η_7	S	BOR	Shortages	Unconditional	Conditional
0.75	0	0	1	1	0	2	1	5	0.74	0.02	0.07	2.92
0.80	0	0	0	2	0	3	4	9	0.80	0.06	0.18	3.20
0.85	0	0	2	1	2	3	5	13	0.85	0.12	0.42	3.60
0.90	0	1	2	3	2	3	5	16	0.89	0.19	0.74	3.96
0.95	2	1	3	3	2	4	5	20	0.95	0.31	1.41	4.54

5. Discussions and extensions

In this section, we compare our model against the traditional queueing model. We also compare the performance of our model against a traditional optimization approach based on the BOR. As Theorem 3 illustrates, the BOR may overestimate or underestimate the risk depending on the situation. We thus also provide a numerical example showing that optimizing shortages using the BOR may significantly underestimate the risk. Finally, we illustrate an extension of our model to the matter of capacity planning across hospitals, where the rate of emergency arrivals $\boldsymbol{\lambda}$ is a decision variable.

5.1. Comparison with queueing models

Our model allows the admission of patients even where there are bed shortages, which is in fact an infinite capacity offered load model. Instead, in a traditional queueing model, the capacity is fixed

and new patients would wait in the queue in the absence of vacant beds. In this section, we compare the performance of these two models using simulations, and show that the offered load model may be a better representation of inpatient dynamics. For the offered load model, the corresponding set of performance measures would be the probability of bed shortages and expected shortages. For the queueing model, performance measures are the probability of delay or expected queue length.

In Table 5, we varied the number of weekly elective admissions S from 7 to 20. Under the same optimal scheduling policy for elective patients (based on Table 7), we simulated the performance for both the offered load model and the traditional queueing model. Observe that the average queue length in the queueing model is significantly larger than the expected bed shortages in the offered load model. While this may be consistent with queueing theory, queues of such lengths are rarely seen in practice. Indeed, flexible capacity is often created when the present BOR is very high by adding temporary beds or overflowing patients to other wards. This example illustrates that our paradigm may be more applicable than the queueing model when modelling the actual dynamics amongst arrivals and ward congestion.

Table 5 Simulated performance measures of the offered load model and the queueing model

		Offered load model		Queueing model	
S	BOR	Prob. of Shortages	Expected Shortages	Prob. of Delays	Expected Queue Length
7	0.80	0.06	0.20	0.08	0.44
8	0.82	0.08	0.26	0.09	0.53
9	0.83	0.09	0.31	0.13	0.82
10	0.85	0.11	0.39	0.16	1.13
11	0.86	0.13	0.48	0.18	1.37
12	0.87	0.15	0.58	0.24	2.09
13	0.89	0.18	0.70	0.25	2.59
14	0.90	0.20	0.84	0.35	4.03
15	0.92	0.24	1.00	0.41	5.59
16	0.93	0.27	1.18	0.44	7.60
17	0.94	0.30	1.37	0.50	10.98
18	0.96	0.34	1.59	0.65	18.95
19	0.97	0.38	1.84	0.76	33.85
20	0.98	0.41	2.12	0.85	75.73

5.2. Comparison with optimization using BOR

We also compare our throughput-maximizing model which restricts the BSI under some bound ϵ against a traditional optimization model which caps the BOR under ϵ'. The latter results in the model:

$$\begin{aligned} \max \ & \sum_{t\in\mathcal{T}} \eta_t \\ \text{s.t. } & \mathbb{E}[\tilde{n}_t(\boldsymbol{w})] \le \epsilon'\kappa, \ \ \forall t \in \mathcal{T} \\ & \boldsymbol{w} \in \mathcal{W}. \end{aligned} \tag{19}$$

In this comparison, the capacity $\kappa = 44$ is fixed across the planning horizon. When the bounds on the BOR and BSI were both set at $\epsilon = \epsilon' = 0.85$, we find that about 11.11% more elective patients could be admitted from our throughput-maximizing model (16) as compared to the traditional BOR model (19). This makes sense because the BSI is risk-adjusted and Theorem 3 guarantees that it is less than the BOR under the assumption of traditional Poisson arrivals for emergency patients and deterministic arrivals for elective patients. (In Section 5.3, we show a circumstance where the BSI is larger than BOR.) Hence, given the same restrictions, the BSI admits more elective patients (see Table 6). Table 11 shows the results when we increase the capacity from 44 to 60 such that more elective patients could be admitted given the same emergency arrival. We can see that the average BOR under optimization using BSI is larger than that under optimization using BOR (consistent with Theorem 3). The managerial insight for practitioners is that, when the proportion of elective patients is not small, the manager can be more aggressive and set a higher BOR target at 90% rather than 85%. Since elective admissions could be scheduled and contribute less to bed shortages than emergency patients, they could be used to smooth the arrival process and mitigate the risk of bed shortages.

5.3. A numerical study with arrival surges

We consider the case where there is a possible surge in arrival, *e.g.* due to a large-scale accident. Let $\bar{p} = 0.05$ be the probability of the surge and $\Delta = 10$ be the surge arrival rate. This is the case in Corollary 2b, where arrivals are modelled by the doubly stochastic Poisson process with rate $\tilde{\lambda}_t = \lambda_t + \tilde{d}_t\Delta$, and $\tilde{d}_t$ is i.i.d. Bernoulli with probability $\bar{p}$. Recall that λ_t is the usual arrival rate. Under this circumstance, Theorem 3 implies that the BSI is generally larger than BOR, *i.e.*, it is more sensitive to the risk of surges than BOR. When using BOR to plan the weekly elective admissions, it is akin to modelling arrivals as a Poisson process with rate $\lambda_t + \bar{p}\Delta$.

We compare the consequence of using the BOR versus the BSI to optimize under such a situation. Figure 8 shows the optimal number of weekly elective admissions under different metrics. When

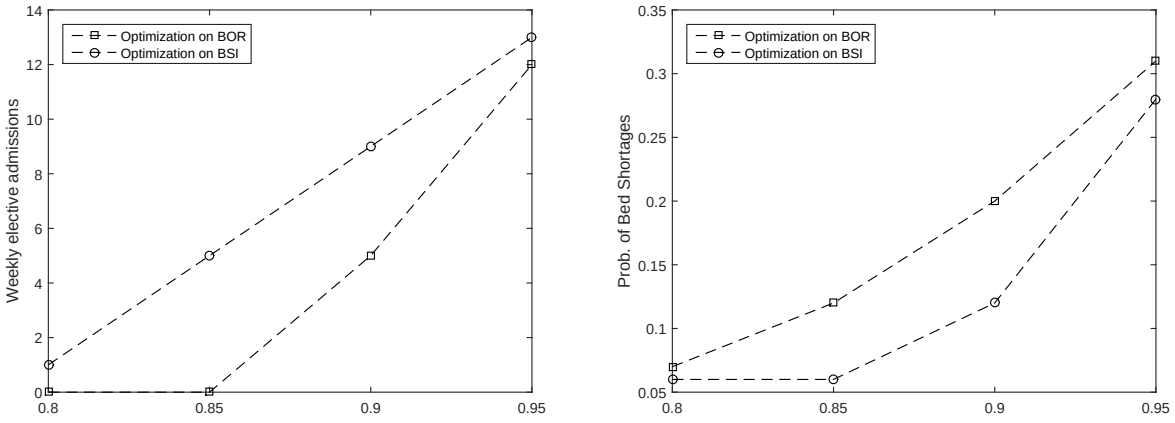
Table 6 The difference between optimization on BOR and BSI with ward capacity 44

	Optimal elective scheduling								Ave.	Prob. of	Expected Shortages	
ϵ	η_1	η_2	η_3	η_4	η_5	η_6	η_7	S	BOR	Shortages	Unconditional	Conditional
Optimization on BOR												
0.75	0	0	0	1	0	0	0	1	0.72	0.02	0.05	2.87
0.80	0	0	1	0	0	2	2	5	0.78	0.04	0.13	3.13
0.85	0	0	0	2	0	3	4	9	0.83	0.09	0.32	3.49
0.90	1	0	2	2	1	3	4	13	0.89	0.18	0.70	3.95
0.95	1	1	2	3	2	3	5	17	0.94	0.30	1.38	4.55
Optimization on BSI												
0.75	0	0	0	1	0	1	0	2	0.73	0.02	0.06	2.93
0.80	0	0	0	1	0	3	2	6	0.79	0.05	0.16	3.20
0.85	0	0	1	2	1	0	6	10	0.85	0.11	0.40	3.62
0.90	1	0	2	2	1	3	5	14	0.90	0.21	0.84	4.09
0.95	1	1	2	3	2	3	5	17	0.94	0.30	1.37	4.55

using BOR to plan, the weekly elective quota increases from 1 to 13 when the BOR threshold increases from 0.8 to 0.95. However, when using BSI to plan, the weekly elective quota is less than the quota using BOR, which is consistent with our theoretical results. Figure 8 also shows that probability of bed shortages is much lower when planing with the BSI. If the manager is risk averse, he will try to set a lower BOR threshold like 0.8 or 0.85 to avoid bed shortages, in which case, the BSI would suggest avoiding admitting elective patients during the periods of surges. The managerial insight gleaned from this numerical example is that, whenever there is possible surge in arrivals or if arrivals are highly unpredictable (for example, during potential epidemic seasons), the manager should be more conservative and admit less elective patients by reserving more beds for these potential arrivals.

5.4. An extension to capacity planning at the regional level - ambulance zoning

In some countries or localities, the total capacity of an entire health system, comprising multiple healthcare facilities with varying roles and specialisations, is planned as a whole (see for example accountable care organizations or ACOs Pham et al. 2015). In Singapore, the network of restructured and community hospitals is partitioned into 3 Regional Health Systems (RHS), where some manpower and resourcing decisions are made centrally for each RHS. These decisions may include:



(a) Elective Admissions (b) Prob. of Shortages

Figure 8 Optimization on BOR versus BSI with burst arrivals

- Zoning: How large should the service population of each hospital be? Where do we draw the geographical lines between the service catchment of each hospital?
- Capacity: How large should the capacity of each hospital be, to sufficiently service its locality?

We examine the problem of ambulance zoning with respect to capacity planning, *i.e.*, to assign the population within each ACO (or RHS in Singapore) to the hospitals, such that the risk of bed shortages in each hospital is minimized.

We present how model (18) could be applied in this context. Let $\mathcal{I}$ denote the set of hospitals. We are not interested in the differentiation of wards in each hospital; we consider the hospitals as a whole. Each hospital has a fixed capacity $\kappa_t^i = \kappa^i$ for all $t \in \mathcal{T}$. Let us assume that for each individual in this population, on a particular day, the risk of admission into ED is the same. Then the expected number of emergency admissions to hospital i on day t is given by λ_t^i and sums to the expected number of admissions across all hospitals, denoted Λ. In this model, each hospital is allowed to have a different LOS survival probability p_t^i and q_t^i, which could arise due to different patient management practices.

Under this interpretation, $\boldsymbol{\lambda}$ enters into the decision variables $\boldsymbol{w}$. $\mathcal{W}$ can be the feasible set imbued with the following constraints for model (18):

- *Population constraint*: $\sum_{i\in\mathcal{I}} \lambda_t^i = \Lambda$.
- *Total capacity cost constraint*: Capacity can only be as large as it is cost effective, $\sum_{i\in\mathcal{I}} m^i \kappa^i \leq C$, where m^i is the cost vector for each hospital.
- *Service continuity constraint*: Serves a minimal number of elective patients, $\sum_{t\in\mathcal{T}} \eta_t^i \geq S^i$, $\forall i \in \mathcal{I}$.

6. Conclusion

Beds are a critical resource in a hospital and we can apply our proposed analytics framework for their effective management, which could reduce costs and improve the quality of care of patients. The generality of our proposed framework will allow all analytics to be housed under one system, thus further saving costs. In the context of prescriptive analytics, our approach of optimization via BSI leads to highly tractable and scalable linear optimization problems, which permits us to effectively incorporate large sets of data with the precision of patients characterized by type and demographic, each with different arrivals and LOS profiles. Although we have focused on steady-state models, the framework can be extended to dynamic settings for managing a collection of healthcare units where the risks bed shortages are mitigated holistically across the entire system.

References

Akushevich, I., J. Kravchenko, S. Ukraintseva, K. Arbeev, and A. Yashin (2013). Time trends of incidence of age-associated diseases in the us elderly population: Medicare-based analysis. *Age Ageing 42*(4), 494–500.

Anderson, D., C. Price, B. Golden, W. Jank, and E. Wasil (2011). Examining the discharge practices of surgeons at a large medical center. *Health care management science 14*(4), 338–347.

Andreoni, J. and J. Miller (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica 70*(2), 737–753.

Aumann, R. J. and R. Serrano (2008). An economic index of riskiness. *Journal of Political Economy 116*(5), 810–836.

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer (2004). Modeling daily arrivals to a telephone call center. *Management Science 50*(7), 896–908.

Ben-Tal, A. and M. Teboulle (2007). An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance 17*(3), 449–476.

Biber, R., H. Bail, C. Sieber, P. Weis, M. Christ, and K. Singler (2013). Correlation between age, emergency department length of stay and hospital admission rate in emergency department patients aged $\geq$ 70 years. *Gerontology 59*(1), 17–22.

Brown, D. B. and M. Sim (2009). Satisficing measures for analysis of risky positions. *Management Science 55*(1), 71–84.

Buck, R. (1958). Preferred optimal strategies. *Proceedings of the American Mathematical Society 9*(2), 312–314.

Challen, K., A. Bentley, J. Bright, and D. Walter (2007). Clinical review: Mass casualty triage – pandemic influenza and critical care. *Crit Care 11*(2), 212.

Chan, C. W., V. F. Farias, and G. J. Escobar (2016). The impact of delays on service times in the intensive care unit. *Management Science 63*(7), 2049–2072.

Clements, A., K. Halton, N. Graves, A. Pettitt, A. Morton, D. Looke, and M. Whitby (2008). Overcrowding and understaffing in modern health-care systems: key determinants in meticillin-resistant staphylococcus aureus transmission. *Lancet Infect Dis 8*(7), 427–34.

Cochran, J. K. and K. Roche (2008). A queuing-based decision support methodology to estimate hospital inpatient bed demand. *Journal of the Operational Research Society 59*(11), 1471–1482.

Dai, J. and P. Shi (2018). Inpatient bed overflow: An approximate dynamic programming approach. *Manufacturing and Service Operations Management, Forthcoming*.

Esogbue, A. O. and A. J. Singh (1976). A stochastic model for an optimal priority bed distribution problem in a hospital ward. *Operations Research 24*(5), 884–898.

Föllmer, H. and A. Schied (2002, Oct). Convex measures of risk and trading constraints. *Finance and Stochastics 6*(4), 429–447.

Forster, A. J., I. Stiell, G. Wells, A. J. Lee, and C. van Walraven (2003). The effect of hospital occupancy on emergency department length of stay and patient disposition. *Acad Emerg Med 10*(2), 127–33.

Frittelli, M. and E. R. Gianin (2005). Law invariant convex risk measures. In *Advances in mathematical economics*, pp. 33–46. Springer.

Green, L. V. (2002). How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing 39*(4), 400–412.

Gupta, D. and B. Denton (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE transactions 40*(9), 800–819.

Hall, N. G., D. Z. Long, J. Qi, and M. Sim (2015). Managing underperformance risk in project portfolio selection. *Operations Research 63*(3), 660–675.

Haseltine, W. A. (2013). *Affordable Excellence: The Singapore Healthcare Story*. Brookings Institution Press Washington, D.C.

Helm, J. E. and M. P. Van Oyen (2014). Design and optimization methods for elective hospital admissions. *Operations Research 62*(6), 1265–1282.

Hershey, J. C. and P. J. Schoemaker (1985). Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science 31*(10), 1213–1231.

Kaier, K., N. T. Mutters, and U. Frank (2012). Bed occupancy rates and hospital-acquired infections–should beds be kept empty? *Clin Microbiol Infect 18*(10), 941–5.

Kao, E. P. C. and G. G. Tung (1981). Bed allocation in a public health care delivery system. *Management Science 27*(5), 507–520.

KC, D. S. and C. Terwiesch (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science 55*(9), 1486–1498.

KC, D. S. and C. Terwiesch (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management 14*(1), 50–65.

Khachiyan, L. (1989). The problem of calculating the volume of a polyhedron is enumerably hard. *Russian Mathematical Surveys 44*(3), 199–200.

Kim, S.-H., P. Vel, W. Whitt, and W. C. Cha (2015). Poisson and non-Poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters 43*(3), 247–253.

Kim, S.-H. and W. Whitt (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management 16*(3), 464–480.

Kuntz, L., S. Scholtes, and A. Vera (2007). Incorporating efficiency in hospital-capacity planning in germany. *The European Journal of Health Economics 8*(3), 213–223.

Lamiri, M., X. Xie, A. Dolgui, and F. Grimaud (2008). A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research 185*(3), 1026–1037.

Lynn, J., B. M. Straube, K. M. Bell, S. F. Jencks, and R. T. Kambic (2007). Using population segmentation to provide better health care for all: the "Bridges to Health" model. *Milbank Q 85*(2), 185–208; discussion 209–12.

Madsen, F., S. Ladelund, and A. Linneberg (2014). High levels of bed occupancy associated with increased inpatient and thirty-day hospital mortality in Denmark. *Health Affairs 33*(7), 1236–1244.

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

Nemirovski, A. and A. Shapiro (2006). Convex approximations of chance constrained programs. *SIAM Journal on Optimization 17*(4), 969–996.

Pham, H. H., J. Pilotte, R. Rajkumar, E. Richter, S. Cavanaugh, and P. H. Conway (2015). Medicare's vision for delivery-system reform — the role of ACOs. *New England Journal of Medicine 373*(11), 987–990. PMID: 26352812.

Qi, J. (2017). Mitigating delays and unfairness in appointment systems. *Management Science 63*(2), 566–583.

Rockafellar, R. T., S. Uryasev, et al. (2000). Optimization of conditional value-at-risk. *Journal of risk 2*, 21–42.

Rose, M., H. Pan, M. R. Levinson, and M. Staples (2014). Can frailty predict complicated care needs and length of stay? *Intern Med J 44*(8), 800–5.

Schull, M. J., J. P. Szalai, B. Schwartz, and D. A. Redelmeier (2001). Emergency department overcrowding following systematic hospital restructuring: trends at twenty hospitals over ten years. *Acad Emerg Med 8*(11), 1037–43.

Shi, P., M. C. Chou, J. G. Dai, D. Ding, and J. Sim (2016). Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science 62*(1), 1–28.

Song, H., A. Tucker, R. Graue, S. Moravick, and J. Yang (2018). Capacity pooling in hospitals: The hidden consequences of off-service placement.

Sprivulis, P. C., J. A. Da Silva, I. G. Jacobs, A. R. Frazer, and G. A. Jelinek (2006). The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Med J Aust 184*(5), 208–12.

Teow, K. L., E. El-Darzi, C. Foo, X. Jin, and J. Sim (2012). Intelligent analysis of acute bed overflow in a tertiary hospital in Singapore. *J Med Syst 36*(3), 1873–82.

Virtanen, M., J. Pentti, J. Vahtera, J. E. Ferrie, S. A. Stansfeld, H. Helenius, M. Elovainio, T. Honkonen, K. Terho, T. Oksanen, and M. Kivimaki (2008). Overcrowding in hospital wards as a predictor of antidepressant treatment among hospital staff. *Am J Psychiatry 165*(11), 1482–6.

Waltz, F. (1967). An engineering approach: Hierarchical optimization criteria. *IEEE Transactions on Automatic Control 12*(2), 179–180.

Whitt, W. (1999). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters 24*(5), 205–212.

Zeldes, S. P. (1989). Optimal consumption with stochastic income: Deviations from certainty equivalence. *The Quarterly Journal of Economics 104*(2), 275–298.

Zhou, Y. Y., W. Wong, and H. Li (2014). Improving care for older adults: a model to segment the senior population. *Perm J 18*(3), 18–21.

Notes

[1] Khalik, S. Public hospitals 'borrowing' ward space, The Straits Times, 30 Aug 2011. Article retrieved from `https://www.ttsh.com.sg/about-us/newsroom/news/article.aspx?id=2422`.

[2] Parliamentary query raised to the Minister of Health, 20 Jan 2014. Reply retrieved from `https://www.moh.gov.sg/content/moh_web/home/pressRoom/Parliamentary_QA/2014/bed-crunch.html`.

APPENDICES

A. Proof of Theorem 1

Notice that given a function $\Phi : [0,\infty] \to [0,1]$ as described in the assumptions, $\beta : \mathcal{Z} \to [0,1]$ is a risk-adjusted BOR if and only if $\beta(\tilde{z}) = \Phi(\rho(\tilde{z}))$ and $\rho : \mathcal{Z} \to [0,\infty]$ is lower semi-continuous with the properties: For all $\tilde{z}, \tilde{z}_1, \tilde{z}_2 \in \mathcal{Z}$,

(a) **Monotonicity**: If $\mathbb{P}[\tilde{z}_1 \leq \tilde{z}_2] = 1$, then $\rho(\tilde{z}_1) \leq \rho(\tilde{z}_2)$.

(b) **Risk-free** If $\mathbb{P}[\tilde{z} \leq 0] = 1$, then $\rho(\tilde{z}) = 0$.

(c) **Overloading**: If $\mathbb{E}[\tilde{z}] > 0$, then $\rho(\tilde{z}) = \infty$.

(d) **Quasiconvexity**: For all $\lambda \in [0,1]$, $\rho(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2) \leq \max\{\rho(\tilde{z}_1), \rho(\tilde{z}_2)\}$.

This is easily true, since all the properties of β simply translate to ρ via the one-to-one function Φ.

We next show that if μ_α is as defined, then $\rho(\tilde{z}) = \inf\{\alpha > 0 \mid \mu_\alpha(\tilde{z}) \leq 0\}$ satisfies the above properties. For this section of the proof, denote by $M(\tilde{z}) = \{\alpha > 0 \mid \mu_\alpha(\tilde{z}) \leq 0\}$.

(a) If $\mathbb{P}[\tilde{z}_1 \leq \tilde{z}_2] = 1$, then for all $\alpha > 0$, $\mu_\alpha(\tilde{z}_1) \leq \mu_\alpha(\tilde{z}_2)$. Thus if $\alpha^* \in M(\tilde{z}_2)$, then $\alpha^* \in M(\tilde{z}_1)$, *i.e.*, $M(\tilde{z}_2) \subseteq M(\tilde{z}_1)$. Thus, $\rho(\tilde{z}_1) = \inf M(\tilde{z}_2) \leq \inf M(\tilde{z}_2) = \rho(\tilde{z}_2)$.

(b) From the above, in particular, if $\tilde{z}_2 = 0$ is the 0 random variable, then $\mu_\alpha(\tilde{z}_1/\alpha) \leq \mu_\alpha(0) = 0$, $\forall \alpha > 0$, thus $M(\tilde{z}_1) = \mathbb{R}^+$ and so $\rho(\tilde{z}_1) = 0$.

(c) If $\mathbb{E}[\tilde{z}] > 0$, then certainly $\forall \alpha > 0$, $\mu_\alpha(\tilde{z}) > 0$. Hence $M(\tilde{z}) = \emptyset$ and so $\rho(\tilde{z}) = \infty$.

(d) We first show that if $\alpha \in M(\tilde{z}_1) \cap M(\tilde{z}_2)$, then $\alpha \in M(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2)$. Indeed, if $\alpha \in M(\tilde{z}_1), M(\tilde{z}_2)$, then $\mu_\alpha(\tilde{z}_1), \mu_\alpha(\tilde{z}_2) \leq 0$. By quasiconvexity of μ_α, we get that $\mu_\alpha(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2) \leq \max\{\mu_\alpha(\tilde{z}_1), \mu_\alpha(\tilde{z}_2)\} \leq 0$. Thus, $\rho(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2) \leq \inf M(\tilde{z}_1) \cap M(\tilde{z}_2) \overset{(*)}{\leq} \max\{\inf M(\tilde{z}_1), \inf M(\tilde{z}_2)\} = \max\{\rho(\tilde{z}_1), \rho(\tilde{z}_2)\}$, where inequality (*) occurs due to lower semi-continuity of μ_α.

The last detail we have to show is that ρ is lower semi-continuous. For all subsequent proofs regarding continuity, we use the fact that a set $\mathcal{Z}$ is closed if and only if for every sequence $\tilde{z}_n \to \tilde{z}$ in the sup-norm topology in $\mathcal{Z}$, its limit $\tilde{z} \in \mathcal{Z}$. In the case of lower semi-continuity, whenever $\rho(\tilde{z}_n) \leq a$, we have $\rho(\tilde{z}) \leq a$.

Fix $a \geq 0$. Consider a sequence $\tilde{z}_n \to \tilde{z}$ such that $\rho(\tilde{z}_n) \leq a$, $\forall n$. We need to show that $\rho(\tilde{z}) \leq a$. Consider two cases: First, suppose $a > 0$. Then since $\rho(\tilde{z}_n) \leq a$, there must exist, for each n, a sequence $a_m^{(n)} \searrow a$ such that $\mu_{a_m^{(n)}}(\tilde{z}_n) \leq 0$, $\forall m$. By the fact that μ_α is lower semi-continuous in α, $\mu_a(\tilde{z}_n) \leq 0$, $\forall n$. But now, as $n \to \infty$, by lower semi-continuity of μ_α, we have that $\mu_a(\tilde{z}) \leq 0$. Hence, $\rho(\tilde{z}) \leq a$ taking infimums.

In the second case, suppose that $a = 0$. Since ρ is non-negative by definition, $\rho(\tilde{z}_n) = 0$ and we strive to prove that $\rho(\tilde{z}) = 0$. But $\rho(\tilde{z}_n) = 0$ implies that $\mu_\alpha(\tilde{z}_n) \leq 0$, $\forall \alpha > 0$. Hence for each $\alpha > 0$, as $n \to \infty$, by lower semi-continuity of μ_α, $\mu_\alpha(\tilde{z}) \leq 0$. But this just means that $\rho(\tilde{z}) = 0$.

Conversely, we want to show that if $\rho(\tilde{z})$ satisfies the above properties, then there exists some functional μ_α as in Theorem 1 and $\rho(\tilde{z}) = \inf\{\alpha > 0 \mid \mu_\alpha(\tilde{z}) \leq 0\}$.

Let $\mu_\alpha(\tilde{z}) = \inf\{v \in \mathbb{R} \mid \rho(\tilde{z} - v) \leq \alpha\}$, where ρ obeys the above properties. From this definition, clearly μ_α is non-increasing in α. We first prove that this choice of μ_α satisfies its required properties. Let $R_\alpha(\tilde{z}) = \{v \in \mathbb{R} \mid \rho(\tilde{z} - v) \leq \alpha\}$.

(a) Suppose $\mathbb{P}[\tilde{z}_1 \leq \tilde{z}_2] = 1$, hence $\mathbb{P}[\tilde{z}_1 - v \leq \tilde{z}_2 - v] = 1$, $\forall v \in \mathbb{R}$. Thus, by monotonicity of ρ, $\rho(\tilde{z}_1 - v) \leq \rho(\tilde{z}_2 - v)$. Hence, if $v \in R_\alpha(\tilde{z}_2)$, then $v \in R_\alpha(\tilde{z}_1)$, or in other words, $\mu_\alpha(\tilde{z}_1) \leq \mu_\alpha(\tilde{z}_2)$.

(b) Let $\mathbb{E}[\tilde{z}] > 0$. Fix $\alpha > 0$. Then if $v < \mathbb{E}[\tilde{z}]$, then $\rho(\tilde{z} - v) = \infty$ by overloading. Thus, $v \notin R_\alpha(\tilde{z})$. This means that $\mu_\alpha(\tilde{z}) \geq \mathbb{E}[\tilde{z}] > 0$.

(c) Consider $R_\alpha(0)$. Now $\rho(-v\alpha) = \infty$ for all $v < 0$ by overloading avoidance. So 0 is a lower bound to $R_\alpha(0)$. But $0 \in R_\alpha(0)$ by satisficing. Hence $\mu_\alpha(0) = 0$.

(d) Let $v \in R_\alpha(\tilde{z}_1) \cap R_\alpha(\tilde{z}_2)$. Hence, $\rho(\tilde{z}_i - v) \leq \alpha$, $i = 1, 2$. By quasiconvexity, $\rho(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2 - v) \leq \max\{\rho(\tilde{z}_1 - v), \rho(\tilde{z}_2 - v)\} \leq \alpha$. Hence, $v \in R_\alpha(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2)$, which means that $\mu_\alpha(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2) \leq \inf R_\alpha(\tilde{z}_1) \cap R_\alpha(\tilde{z}_2) \overset{(*)}{\leq} \max\{\inf R_\alpha(\tilde{z}_1), \inf R_\alpha(\tilde{z}_2)\} = \max\{\mu_\alpha(\tilde{z}_1), \mu_\alpha(\tilde{z}_2)\}$. Again, the (*) inequality holds by lower semi-continuity of ρ.

We now prove that ρ can indeed be constructed from this μ_α. Let $\rho'(\tilde{z}) = \inf\{\alpha' > 0 \mid \mu_{\alpha'}(\tilde{z}) \leq 0\}$. We show that $\rho' = \rho$. Consider $\alpha > 0$ such that $\mu_\alpha(\tilde{z}) \leq 0$, *i.e.*, $\inf\{v \in \mathbb{R} \mid \rho(\tilde{z} - v) \leq \alpha\} \leq 0$. Notice that $\forall \alpha \geq \rho(\tilde{z})$, this is satisfied since $v = 0$ clearly lies in this set. Hence, $\rho(\tilde{z})$ lies in the set of α's such that $\inf\{v \in \mathbb{R} \mid \rho(\tilde{z} - v) \leq \alpha\} \leq 0$. As such, $\rho'(\tilde{z}) \leq \rho(\tilde{z})$.

To prove equality, suppose that instead there exists some $\tilde{z}$ where $\rho'(\tilde{z}) < \rho(\tilde{z})$, or in other words, there is some $\alpha < \rho(\tilde{z})$ such that $\inf\{v \in \mathbb{R} \mid \rho(\tilde{z} - v) \leq \alpha\} \leq 0$. Denote by E, the set $\{v \in \mathbb{R} \mid \rho(\tilde{z} - v) \leq \alpha\}$. We first assert that 0 is a lower bound for E. Consider $v \leq 0$. Then by monotonicity of ρ, $\mathbb{P}[\tilde{z} \leq \tilde{z} - v] = 1$ implies that $\alpha < \rho(\tilde{z}) \leq \rho(\tilde{z} - v)$. Thus, $v \notin E$, $\forall v \leq 0$. Thus 0 is a lower bound for E.

But by assumption, $\inf E \leq 0$. Hence, $\inf E = 0$ but where $0 \notin E$. Thus there exists a sequence $v_n \searrow 0$ such that $\rho(\tilde{z} - v_n) \leq \alpha$. But by lower semi-continuity of ρ, $\tilde{z} - v_n \to \tilde{z}$ as $n \to \infty$, so $\rho(\tilde{z}) \leq \alpha < \rho(\tilde{z})$, which is a clear contradiction.

Finally, it leaves to show that μ_α is lower semi-continuous for every $\alpha > 0$ and also in α. We do this only for the former as the proof for the latter is identical. Fix $\alpha > 0$. For each $v \in \mathbb{R}$, take a sequence $\tilde{z}_n \to \tilde{z}$ such that $\mu_\alpha(\tilde{z}_n) \leq v$. Now, for each n, there exists a sequence $v_m^{(n)} \to v$ such that $\rho(\tilde{z}_n - v_m^{(n)}) \leq \alpha$. Using the lower semi-continuity of ρ, we have that for each n, $\tilde{z}_n - v_m^{(n)} \to \tilde{z} - v$, hence, $\rho(\tilde{z}_n - v) \leq \alpha$, $\forall n$. But applying lower semi-continuity again on $\tilde{z}_n - v \to \tilde{z} - v$, we have that $\rho(\tilde{z} - v) \leq \alpha$. In particular, this means that μ_α, being the infimum, satisfies $\mu_\alpha(\tilde{z}) \leq v$. □

B. Discussion on Lexicographic Minimization

B.1. Justification of the lex min model

Lexicographic minimization was most initially mooted in the game theory context (such as Buck 1958), and methods were soon developed to implement this (as in Waltz 1967). Define the lexicographic order on $\mathbb{R}^N_+$ as follows: For $\mathbf{u} \in \mathbb{R}^N_+$, let $(u_{(1)}, \ldots, u_{(N)})$ define the ordered statistic of $\mathbf{u}$, $u_{(1)} \geq \ldots \geq u_{(N)}$. We say that $\mathbf{u}$ is strictly lexicographically dominated by $\mathbf{v}$, denoted $\mathbf{u} \prec_{\text{lex}} \mathbf{v}$, if $\exists n$ such that $u_{(i)} = v_{(i)}$, $\forall i \in \{1, \ldots, n-1\}$ (possibly empty) and $u_{(n)} < v_{(n)}$. Furthermore, we say that $\mathbf{u}$ and $\mathbf{v}$ are lexicographically equal, denoted $\mathbf{u} =_{\text{lex}} \mathbf{v}$, if $u_{(i)} = v_{(i)}$, for $i = 1, \ldots, N$. We say that $\mathbf{v}$ lexicographically dominates $\mathbf{u}$, written $\mathbf{u} \preceq_{\text{lex}} \mathbf{v}$ if $\mathbf{u} \prec_{\text{lex}} \mathbf{v}$ or $\mathbf{u} =_{\text{lex}} \mathbf{v}$. This now defines a partial order on $\mathbb{R}^N_+$.

We wish to perform a lex min on $(\beta(\tilde{z}^i_t(\boldsymbol{w})))_{t\in\mathcal{T}, i\in\mathcal{I}}$. This approach is desirable in the sense that it represents a degree of 'fairness' across different wards i and time periods t. More specifically, the optimal solution occurs where the risk $\beta(\tilde{z}^i_t)$ is distributed according to how much risk can be afforded to be borne by each time period t and ward i. Given two feasible decisions $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, and their resultant BSIs $\boldsymbol{\beta}_1 = (\beta(\tilde{z}^i_t(\boldsymbol{w}_1)))_{t\in\mathcal{T}, i\in\mathcal{I}}$ and $\boldsymbol{\beta}_2 = (\beta(\tilde{z}^i_t(\boldsymbol{w}_2)))_{t\in\mathcal{T}, i\in\mathcal{I}}$, if $\boldsymbol{\beta}_1 \preceq_{\text{lex}} \boldsymbol{\beta}_2$, then $\boldsymbol{\beta}_1$ is more desirable. Indeed, suppose $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ differ first at $(\bar{t}, \bar{i})$. Then, we *can afford* to lower the metric at $(\bar{t}, \bar{i})$ for $\boldsymbol{\beta}_2$ without causing any ward that is currently better off than $\beta(\tilde{z}^{\bar{i}}_{\bar{t}}(\boldsymbol{w}_2))$, worse than it after the change.

B.2. Algorithm for solving model (18)

Assuming that the system is stable, *i.e.*, there exists $\boldsymbol{w} \in \mathcal{W}$ such that $\mathbb{E}[\tilde{z}^i_t(\boldsymbol{w})] \leq 0$ for all $(t, i) \in \mathcal{T} \times \mathcal{I}$, then this lex min model may be implemented as described in Algorithm 1.

We explain the algorithm here. Let ρ be the riskiness index corresponding to the BSI. The sets $\mathcal{P}$ and $\mathcal{U}$ represent the sets of *processed* and *unprocessed* indices respectively. Observe that $V(\gamma)$ is nondecreasing in $\gamma \geq 0$ and if $\mathcal{W}$ is polyhedral, its solution can be obtained by solving a linear optimization problem. Hence, using bisection search, we can tractably determine γ^* such that $V(\gamma^*) = 0$. Moreover, because l is minimized at 0, there must exist some constraints for which $\rho(\tilde{z}^{\hat{i}}_{\hat{t}}(\boldsymbol{w}^*)) = \gamma^*$. This constraint is the one that generates the next lexicographically largest risk. Hence, we accept its index $(\hat{t}, \hat{i})$ into the processed indices. In the next iteration, since μ_α is nondecreasing in α, if this constraint is replaced with $\mu_{\gamma^*}(\tilde{z}^{\hat{i}}_{\hat{t}}(\boldsymbol{w})) \leq 0$, then the resulting subproblem can only yield $\gamma \leq \gamma^*$.

Algorithm 1 Pseudo-code for Lexicographic Mininimization Model

Initialization: $\mathcal{P} \leftarrow \emptyset$, $\mathcal{U} \leftarrow \{(t,i) \in \mathcal{T} \times \mathcal{I}\}$

Sub-problem: Define sub-problem with inputs $\gamma \geq 0$

$$\begin{aligned} V(\gamma) \triangleq \min\ & l \\ \text{s.t. } & \mu_\gamma(\tilde{z}_t^i(\boldsymbol{w})) \leq l,\ \forall (t,i) \in \mathcal{U} \\ & \mu_\alpha(\tilde{z}_t^i(\boldsymbol{w})) \leq 0,\ \forall (t,i,\alpha) \in \mathcal{P} \\ & \boldsymbol{w} \in \mathcal{W}. \end{aligned}$$

while $\mathcal{U} \neq \emptyset$ **do**

Find γ^* for which $V(\gamma^*) = 0$ and obtain solution $\boldsymbol{w}^*$

$\mathcal{R} \leftarrow \{(t,i) \in \mathcal{U} \mid \rho(\tilde{z}_t^i(\boldsymbol{w}^*)) = \gamma^*\}$

$\mathcal{P} \leftarrow \mathcal{P} \cup \{(t,i,\gamma^*) \mid (t,i) \in \mathcal{R}\}$

$\mathcal{U} \leftarrow \mathcal{U} - \mathcal{R}$

end while

Output: Solution $\boldsymbol{w}^*$

C. Discussion on predictive analytics

We can incorporate the BSI into predictions about bed shortages in the future due to, among other things, demographic drifts towards an aging population. By classifying patients into age groups and estimating their corresponding arrival and survival rates, we may predict how bed shortages will shift. We could also apply the metric in counter-factual analysis such as evaluating the impact of transferring long staying patients from acute hospitals to community hospitals. How it does so boils down to the two uncertainties that shape bed shortages – the arrival and survival rate of patients. The survival rate of patient is defined broadly by the entry event of the patient into the care (sub-)system, such as acute-care hospitals, till the departure event, which can be discharge, step-down care or death.

C.1. Patient segmentation

Let $\mathcal{G}$ be some segmentation of the patients in the ward, perhaps by demographics. To simplify the analysis, we assume that the underlying proportion of each segment $g \in \mathcal{G}$ of elective (respectively, emergency) patients is ϕ_g (respectively, θ_g) and is fixed across all time periods. From the predictive perspective, one can estimate the survival rates p_s^g and q_s^g along these segmentations $g \in \mathcal{G}$. These rates will then affect the eventual demand of beds $\tilde{x}_{t,s}$ and $\tilde{y}_{t,s}$ (see Proposition 1). Statistical techniques to do so have been described in the extant literature and have been illustrated to achieve

a high-level of predictive accuracy over a wide spectrum of segmentation constructs (*e.g.*, Lynn et al. 2007, Zhou et al. 2014). Our goal here is not to discuss them. Instead, we wish to illustrate how predictive analytics may be utilised within the current analysis set-up to address present and future problems.

For simplicity, we assume that the planner does not discern elective patients by specifying admission quotas for each segment, though this can easily be modeled. Mathematically, this is

$$\tilde{x}_{t,s} = \sum_{j=1}^{\eta_t} \tilde{v}_{j,\tilde{g}_j},$$

where $\tilde{v}_{j,\tilde{g}_j}$ represents the Bernoulli distributed outcome of whether patient j (from a random segment $\tilde{g}_j$) scheduled for admission on day t stays at least to the end of day $t+s$. Thus, $\mathbb{P}[\tilde{v}_{j,\tilde{g}_j} = 1 \mid \tilde{g}_j = g] = p_s^g$, and $\mathbb{P}[\tilde{g}_j = g] = \phi_g$. Since $\tilde{v}_{j,\tilde{g}_j}$ are independent and identically distributed,

$$\mathbb{P}[\tilde{v}_{j,\tilde{g}_j} = 1] = \sum_{g\in\mathcal{G}} \phi_g \mathbb{P}[\tilde{v}_{j,\tilde{g}_j} = 1 \mid \tilde{g}_j = g] = \sum_{g\in\mathcal{G}} \phi_g p_s^g \triangleq \bar{p}_s.$$

Hence, $\tilde{x}_{t,s}$ is equivalent to a binomial distribution, where the probability of success is the weighted sum of the survival probabilities, *i.e.*, $\tilde{x}_{t,s} \sim \mathrm{Bin}(\eta_t, \bar{p}_s)$.

The fact that we did not know the segment of each patient is an important assumption here. This makes $\tilde{g}_j$ a random variable and enables the outcomes of each patient to be modelled identically. Indeed, this is not equivalent to modelling $\eta_t^g = \eta_t \phi_g$ and $\tilde{x}_{t,s}^g \sim \mathrm{Bin}(\eta_t^g, p_s^g)$, as $\sum_{g\in\mathcal{G}} \tilde{x}_{t,s}^g \not\sim \mathrm{Bin}(\eta_t, \bar{p}_s)$ in general because patients comprising $\tilde{x}_{t,s}^g$ are not identical across $g \in \mathcal{G}$.

By taking into account demographic drift, we can anticipate how future emergency arrival rate λ_t and segment proportion θ_g might change. Observe that

$$\tilde{y}_{t,s} = \sum_{g\in\mathcal{G}} \tilde{y}_{t,s}^g,$$

where $\tilde{y}_{t,s}^g \sim \mathrm{Pois}(\lambda_t \theta_g q_s^g)$ is the number of emergency patients from segment g who were admitted on day t and has remained in the ward at the end of day $t+s$. Since the sum of independently distributed Poisson random variables is also a Poisson random variable, we have $\tilde{y}_{t,s} \sim \mathrm{Pois}(\lambda_t \bar{q}_s)$, where $\bar{q}_s = \sum_{g\in\mathcal{G}} \theta_g q_s^g$ is the weighted sum of survival probabilities.

C.2. Ageing population and long stayers in hospitals

We zoom in on the challenges of an ageing population – $\mathcal{G}$ would represent age groups here. Some implications of an ageing population on bed shortages are listed as follows:

- Total patient arrivals: Both total emergency arrivals and elective quotas will see higher numbers. The frequency of complex acute and chronic medical conditions can increase with age. With a higher proportion of the population being advanced in age, the total patient arrivals can be expected to increase (Akushevich et al. 2013).

- LOS: Older patients tend to stay longer (*e.g.*, Rose et al. 2014, Biber et al. 2013), due to greater treatment complexity. This lowers the relative treatment efficiency of older patients. With a higher proportion of older patients amongst patient arrivals, the average LOS can be expected to increase.

These effects have an exacerbating effect on bed shortages. In the first instance, one can extrapolate the rate of emergency arrivals from demographic drifts. The same can be done for elective admissions, given other data on patient referrals and treatment. As such, one can examine how the BSI of wards would shift over time or examine the capacity κ_t needed to keep the BSI at today's level.

D. Additional Tables

Table 7 The optimal elective scheduling and the corresponding daily BSI in one week under the risk minimization model when given the weekly elective quota S

	Elective scheduling							Daily BSI						
S	η_1	η_2	η_3	η_4	η_5	η_6	η_7	$\beta(\tilde{z}_1)$	$\beta(\tilde{z}_2)$	$\beta(\tilde{z}_3)$	$\beta(\tilde{z}_4)$	$\beta(\tilde{z}_5)$	$\beta(\tilde{z}_6)$	$\beta(\tilde{z}_7)$
7	0	0	1	1	0	2	3	0.79	0.80	0.80	0.80	0.79	0.79	0.77
8	0	0	1	1	1	2	3	0.80	0.81	0.81	0.80	0.81	0.81	0.79
9	0	0	1	1	1	2	4	0.82	0.83	0.82	0.81	0.82	0.82	0.81
10	0	0	1	2	0	2	5	0.83	0.84	0.83	0.84	0.82	0.82	0.84
11	1	0	1	2	1	2	4	0.85	0.85	0.84	0.85	0.85	0.84	0.83
12	0	0	2	2	1	2	5	0.85	0.86	0.87	0.87	0.86	0.85	0.86
13	1	0	2	2	1	3	4	0.87	0.87	0.88	0.88	0.87	0.88	0.86
14	1	0	2	2	1	3	5	0.89	0.89	0.89	0.89	0.88	0.89	0.89
15	1	0	2	2	2	3	5	0.91	0.90	0.90	0.89	0.91	0.91	0.91
16	1	1	2	2	2	3	5	0.91	0.92	0.92	0.91	0.92	0.92	0.92
17	1	1	2	3	2	3	5	0.92	0.93	0.93	0.94	0.94	0.93	0.93
18	2	1	2	3	2	3	5	0.95	0.95	0.94	0.95	0.95	0.94	0.94
19	2	1	3	3	2	3	5	0.96	0.96	0.97	0.97	0.97	0.96	0.95
20	2	1	3	3	2	3	6	0.99	0.98	0.99	0.98	0.98	0.96	0.99

Table 8 The simulated performance under the risk minimization model when given the weekly elective quota S

	Prob. of Shortage Levels $\sum_{t\in\mathcal{T}} \frac{1}{T}\mathbb{P}[\tilde{z}_t \geq \psi]$									Average	Expected Shortages	
S	$\psi=1$	2	3	4	5	6	7	8	9	BOR	Unconditional	Conditional
7	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0.00	0.00	0.80	0.20	3.28
8	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0.00	0.82	0.26	3.39
9	0.09	0.07	0.05	0.03	0.02	0.02	0.01	0.01	0.00	0.83	0.31	3.47
10	0.11	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0.85	0.39	3.59
11	0.13	0.10	0.07	0.05	0.04	0.03	0.02	0.01	0.01	0.86	0.48	3.70
12	0.15	0.12	0.09	0.07	0.05	0.03	0.02	0.02	0.01	0.87	0.58	3.82
13	0.18	0.14	0.11	0.08	0.06	0.04	0.03	0.02	0.01	0.89	0.70	3.95
14	0.20	0.16	0.13	0.10	0.07	0.05	0.04	0.03	0.02	0.90	0.84	4.08
15	0.24	0.19	0.15	0.12	0.09	0.07	0.05	0.03	0.02	0.92	1.00	4.23
16	0.27	0.22	0.17	0.14	0.11	0.08	0.06	0.04	0.03	0.93	1.18	4.40
17	0.30	0.25	0.20	0.16	0.12	0.09	0.07	0.05	0.04	0.94	1.37	4.55
18	0.34	0.28	0.23	0.18	0.15	0.11	0.09	0.06	0.05	0.96	1.59	4.72
19	0.37	0.32	0.26	0.21	0.17	0.13	0.10	0.08	0.06	0.97	1.84	4.91
20	0.41	0.35	0.30	0.24	0.20	0.16	0.12	0.09	0.07	0.98	2.12	5.11

Table 9 The optimal elective scheduling and the daily BSI under risk minimization model without long-stayers given S

	Elective scheduling							Daily BSI						
S	η_1	η_2	η_3	η_4	η_5	η_6	η_7	$\beta(\tilde{z}_1)$	$\beta(\tilde{z}_2)$	$\beta(\tilde{z}_3)$	$\beta(\tilde{z}_4)$	$\beta(\tilde{z}_5)$	$\beta(\tilde{z}_6)$	$\beta(\tilde{z}_7)$
7	0	0	1	1	0	2	3	0.75	0.77	0.77	0.76	0.76	0.76	0.74
8	0	0	1	1	1	2	3	0.76	0.78	0.78	0.77	0.78	0.78	0.75
9	0	0	1	1	1	2	4	0.78	0.79	0.79	0.78	0.79	0.78	0.78
10	0	0	1	2	0	3	4	0.79	0.80	0.80	0.80	0.79	0.80	0.79
11	0	0	1	2	1	2	5	0.81	0.81	0.81	0.81	0.81	0.80	0.82
12	1	0	1	2	1	3	4	0.83	0.83	0.82	0.82	0.82	0.83	0.82
13	1	0	2	2	1	3	4	0.84	0.84	0.84	0.84	0.84	0.84	0.83
14	1	0	2	2	1	3	5	0.86	0.85	0.86	0.85	0.84	0.85	0.85
15	1	0	2	2	2	3	5	0.87	0.86	0.86	0.86	0.87	0.87	0.87
16	1	1	2	2	2	3	5	0.88	0.89	0.88	0.87	0.88	0.88	0.88
17	1	1	2	3	2	3	5	0.89	0.89	0.89	0.90	0.90	0.89	0.89
18	2	1	2	3	2	3	5	0.92	0.91	0.90	0.91	0.91	0.90	0.90
19	2	1	2	3	2	4	5	0.93	0.93	0.91	0.92	0.92	0.93	0.92
20	2	1	3	3	2	4	5	0.94	0.93	0.94	0.94	0.93	0.95	0.93
21	2	2	3	3	2	4	5	0.95	0.96	0.96	0.96	0.95	0.96	0.94

Table 10 The simulated performance under risk minimization model when given S without long stayers

	Prob. of Shortage Levels $\sum_{t\in\mathcal{T}} \frac{1}{T}\mathbb{P}[\tilde{z}_t \geq \psi]$									Average	Expected Shortages	
S	$\psi=1$	2	3	4	5	6	7	8	9	BOR	Unconditional	Conditional
7	0.04	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.77	0.11	3.05
8	0.04	0.03	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.79	0.14	3.12
9	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0.00	0.00	0.80	0.18	3.20
10	0.07	0.05	0.03	0.02	0.02	0.01	0.01	0.00	0.00	0.81	0.22	3.29
11	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0.00	0.83	0.27	3.39
12	0.10	0.07	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0.84	0.34	3.48
13	0.12	0.09	0.06	0.05	0.03	0.02	0.02	0.01	0.01	0.85	0.41	3.59
14	0.14	0.10	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.87	0.51	3.70
15	0.16	0.12	0.09	0.07	0.05	0.04	0.03	0.02	0.01	0.88	0.62	3.83
16	0.19	0.15	0.11	0.08	0.06	0.05	0.03	0.02	0.02	0.89	0.74	3.96
17	0.21	0.17	0.13	0.10	0.08	0.06	0.04	0.03	0.02	0.91	0.87	4.09
18	0.24	0.20	0.15	0.12	0.09	0.07	0.05	0.04	0.03	0.92	1.04	4.24
19	0.28	0.23	0.18	0.14	0.11	0.08	0.06	0.04	0.03	0.93	1.22	4.39
20	0.31	0.26	0.21	0.16	0.13	0.10	0.07	0.05	0.04	0.95	1.41	4.54
21	0.35	0.29	0.23	0.19	0.15	0.12	0.09	0.07	0.05	0.96	1.63	4.72

Table 11 The difference between optimization on BOR and BSI with ward capacity 60

	Optimal elective scheduling								Ave.	Prob. of	Expected Shortages	
ϵ	η_1	η_2	η_3	η_4	η_5	η_6	η_7	S	BOR	Shortages	Unconditional	Conditional
Optimization on BOR												
0.75	2	2	3	2	3	4	6	22	0.74	0.01	0.02	2.79
0.80	4	2	4	4	3	5	5	27	0.79	0.02	0.07	3.14
0.85	3	3	5	5	4	4	8	32	0.84	0.07	0.24	3.57
0.90	5	3	5	6	5	6	7	37	0.90	0.15	0.62	4.12
0.95	5	5	5	6	6	7	8	42	0.95	0.28	1.36	4.83
Optimization on BSI												
0.75	2	2	3	4	3	5	7	26	0.78	0.02	0.06	3.06
0.80	3	1	4	5	4	5	8	30	0.82	0.05	0.16	3.40
0.85	4	3	5	5	4	6	8	35	0.87	0.11	0.43	3.87
0.90	5	4	6	6	5	6	7	39	0.92	0.20	0.87	4.39
0.95	5	4	6	7	4	8	9	43	0.96	0.31	1.58	5.01