A Model for Inpatient Ward Simulation and Forecasting

Timo Latruwe
Department of Information Technology
Ghent University
Technology Lane, Ghent, Belgium
timo.latruwe@ugent.be

Marlies Van der Wee
Department of Information Technology
Ghent University
Technology Lane, Ghent, Belgium

Joke Devriese
Hict
MeetDistrict, Ghent, Belgium

Pieter Vanleenhove
Hict
MeetDistrict, Ghent, Belgium

Sofie Verbrugge
Department of Information Technology
Ghent University
Technology Lane, Ghent, Belgium

Didier Colle
Department of Information Technology
Ghent University
Technology Lane, Ghent, Belgium

Abstract—Making decisions on capacity requirements in hospital environments has long-term consequences. Sound estimates of expected mid and long-term demand for hospital beds are required to avoid oversupply or deficits. Though in many cases, forecasts of admissions and Length of Stay evolutions for specific pathologies are available, current models rarely model hospital environments on the pathology level, rendering them useless. Additionally, existing models are most often unit type or facility-dependent. This work describes a model to estimate care demand and capacity needs per ward of a generic hospital, in which pathology-specific parameters can be used, and that supports patient stays structurally segmented across multiple wards. The methodology enables an increase in model complexity while explicitly guarding the accuracy of parameter estimates.

I. INTRODUCTION

Hospitals provide highly customized services to their patients. As a consequence, there is limited uniformity in the trajectories of patients through specialisms, wards, and nursing units. Modellers are thus confronted with a highly complex processing system with many plausible paths. Given the diversity of treatment and the low prevalence of unique configurations of care trajectories that follows from that diversity, it becomes difficult to accurately estimate model parameters as soon as a significant level of detail on treatment is included. Hospitals generally do not have the scale such that they and their patients can be modelled in detail without undermining statistical accuracy.

Depending on the angle of the study, modellers choose to aggregate across, or make abstraction of, particular details in order to preserve model estimate accuracy. Often, as in Devapriya et al. [1], Komashie et al. [2] and others [3], [4], details on the patient's pathology are not considered. Though the relevant authors do not explicitly identify parameter estimation accuracy as the reason to make abstraction of the patient pathology, it is inevitable that further dissection of patient groups would lead to these issues. For instance, consider 5000 ICU admissions per year. If further subdivided into groups by pathology, parameter estimates of the probability per patient group to transfer to different downstream destinations would deteriorate to a point where they would be considered unacceptably inaccurate.

In the methodology described by Latruwe et al. [5], the choice as to which details to omit in order to ensure parameter estimation accuracy went to operational aspects of the hospital. Rather than yielding pathology particulars, the researchers chose to reduce the granularity of the description of the trajectory. This methodological preference is motivated primarily by a need to apply forecasted instances of the parameters. Since the forecasts are dependent on pathology information, the model needs to function at the level of the pathology. A major limitation of the authors' method is that it restricts the model's ability to estimate capacity for wards that reflect the wards that are physically present in the hospital. Concretely, it is assumed that patient stays are entirely allocated to a particular ward, which does not always accord with real-world conditions, and abstracts a lot of the complexity of treatment patterns. Data from a large Belgian hospital spanning 2016-2019 shows that 89.7% of inpatient stays are entirely attributable to one specific bed type, meaning that the patient does not spend at least a full day in beds of two different types. Wards are organized per bed type, in this system. Nonetheless, those patients that do require care in multiple bed types represent 30.2% of all inpatient days. Given this disproportionate impact on required capacity, enabling more granular insights into patient paths without losing pathology data could improve models of this type.

November 30th, 2020

II. PROBLEM DESCRIPTION

As introduced, this research builds on and extends the methodology described in Latruwe et al. [5]. The primary objectives are to estimate long-term capacity needs for individual hospitals, in a unit and facility-independent way, taking into account expected evolutions in pathology prevalence, treatment characteristics, and local population attributes. This work

develops the possibility to simulate segmented patient stays, whose parts can be allocated to different wards, allowing more specific definitions of inpatient wards that more accurately reflect their real-world counterpart. Additionally, it integrates extrapolation methods compatible with this methodology, used for forecasting future need.

Individual Belgian hospitals generally treat between 5 000 and 40 000 inpatients annually. When categorizing inpatients for a rather large hospital according to the 3M Service Line (v34), which is a broad classification of pathologies, about half of the service lines contain less than one thousand patients annually. Suppose there are five relevant ward types for a Service Line in which 1000 patients are observed and the modeller intends to model patient paths through multiple wards. Equitable distribution would lead to about 200 patients per ward. Subsequent modelling of transfers requires estimation of transfer probabilities and average Lengths of Stay per stay subcomponent. Transfer probability estimates for subgroups would inevitably exhibit confidence intervals between 10 and 15% in size, assuming significance at 95%. Evidently, this uncertainty accumulates with the uncertainty of the initial patient estimates, LoS estimates, and other factors. In this hypothetical example, a large hospital, broad pathology definitions, and equitable distribution per subgroup are used. In addition, only one transfer per patient stay is assumed, while stays representing 15% of days have at least three subcomponents. In practice, many hospitals are smaller, desire more specific patient group definitions, and do not have uniform distributions of patients across ward types. Consequently, the confidence intervals related to much of the model output would be even larger, and not fit for use. This phenomenom is a manifestation of the more general risk in simulation models to include too much detail. In this case, it inhibits more granular modelling of hospital treatments, required to define ward-stays as prefered.

Decoupling transfer probabilities from specific pathologies and measuring them on ward level would resolve this issue in most cases, but it would obfuscate pathology-dependent transfer rates, and with it, the impact that unequal growth in different pathologies has on ward capacity needs.

In this work, parameter estimation issues in this context are addressed while preserving pathology information. Additionally, a methodology is described to extrapolate admission rates and the average LoS with data constraints. The described methodology is flexible, and can be used for small and large hospitals. It can also accomodate different preferences in the trade-off between pathology granularity and accuracy.

III. Methodology

This section contains three parts. First, pathology clustering is discussed. The pathology clustering method aggregates similar pathologies, or pathologies in which patients have similar attributes, into pathology categories such that groups with sufficient observations arise. Subsequently, model parameter estimation is discussed. Lastly, the assumptions used in the extrapolation methodology are elaborated on.

A. Pathology clustering

The pathology clustering method creates categories of pathologies that are homogeneous along particular lines. The characteristics of patients covered by a pathology category are meant to be as analogous as possible, especially along those dimensions that affect capacity needs or the ability to extrapolate category characteristics consistently, such as the Length of Stay, admissions growth, and evolution of the Length of Stay.

Inpatient stays, whose attributes contribute to the clustering process, are defined as stays in a particular bed type or ward, which are synonyms in the context of this paper. As a consequence, an individual patient's physical stay can consist of multiple stays in this sense, if he or she spent at least one day in multiple bed types. Unless stated otherwise, when stays or attributes of stays are referred to in subsequent sections, this definition is applied.

Clustering is performed on the pathology level. Pathologies have defining attributes, which mutually exclusively distinguish them from each other. In the context of this paper, the APR–DRG-code, Severity of Illness, and bed type are used. The attributes used to quantify similarity are 3M Service Line, APR–DRG, Severity of Illness, and average Length of Stay.

A hierarchical clustering algorithm using Ward's method to measure cluster distance is chosen. Hierarchical clustering allows for more or less arbitrary choice in group sizes, which serves the purposes of this research, where lower limits are placed on group sizes though fine granularity and high homogeneity within individual clusters is desired. In this case, each cluster of pathologies is required to cover at least 170 patients. Additional aggregation beyond 170 patients is cut off unless the difference between the incremental distance between the clusters under consideration as compared to the last round of aggregation, is larger than the equivalent incremental distance observed in the last round as compared to the second-to-last round. Hence, it is ensured that tiny incremental increases in heterogeneity are allowed even when a category already contains 170 admissions.

Transformations on pathology attribute values are performed in the distance function, which calculates the distance between any two pathologies. The attributes Service Line and APR-DRG give rise to binary distances between the two evaluated pathologies. An equal or unequal APR-DRG in two different clusters yields a distance of 0 or 1 respectively in the relevant dimension. Identical Service Lines yield a distance of 0, while different Service Lines yield a distance of a large number M. The large number M is meant to dominate the clustering, in the sense that pathologies from different Service Lines should never be grouped together unless the lower limit of admission numbers is not reached. Further, the number is added to the result of the Euclidean distance calculated based on the other dimensions, such that its influence can be clearly identified in the dendrogram resulting from the clustering. Next, the Severity of Illness is linearly transformed to take a value between 0 and 1 Lastly, a multiple of the logarithm of the

average LoS is used to include relative differences rather than absolute ones and increase the importance of the variable in the clustering.

B. Ward simulation: parameter estimation

As explained in section III-A, stays are defined as the time spent in a particular bed type or ward. The admission rate, LoS distribution, and seasonality factors are attributes related to stays under this definition. Consequently, the admission rate for a pathology category is the amount of observed stays in a ward, matching the APR-DRG and SOI combinations from the set of combinations grouped in the clustering phase, regardless of whether patients visit a ward multiple times. The Length of Stay distribution is analogously derived from the duration spent in the specific ward. Lastly, the methodology to introduce seasonality into the model is adopted from Latruwe et al. [5].

Simulations are performed per ward, which are uniquely related to a set of pathology categories. Daily admissions are sampled from a Poisson distribution with the admission rate, corrected for the seasonality pattern, as the mean [5]. Subsequently, the LoS is applied to each stay, sampled from the empirical frequency distribution for the pathology category. After stochastic daily occupancy instances are generated, the occupancy frequency distribution is derived and used as a basis for capacity estimates associated with particular service levels in terms of deficient capacity-days per year, semester, or other.

C. Parameter extrapolation

Since one of the primary use cases for this model is forecasting required capacity in future years, the default parameters of the model are meant to be replaceable with forecasts for future years. By default, the model implements forecasting methods based on demographic and historical data [5]. Given the narrow definition of stays in this model, it is non-trivial to acquire the appropriate data to perform these extrapolations.

Essentially, the extrapolations require a sample or population of observations that are representative in those dimensions relevant for the extrapolation. Concretely, extrapolations are made for the admission rate and the average LoS, based on demographic conditions and on historical trends. They are examined in the same order.

No Belgian national data sources are available that relate pathology-bed type combinations to age groups. Therefore, it is not practically feasible to directly determine the age distribution present in the admission pools as they are defined in this paper. Alternatively, groups that are, in the relevant respects, representative of the admission pools could be determined and used to sample the age distribution. For instance, if the bed type attribute is discarded, the age distribution of the larger group of patients identified by the APR-DRG and Severity of Illness could be used as a basis for the extrapolation. This would presume that the age distribution of patients related to the different bed types is identical, or that the influence of age on admissions or the average Length of Stay is negligible. Both are decidedly flawed assumptions in the context of this

paper. Several bed types under consideration have an explicit link to age. For instance, Geriatric beds are only accessible to patients older than 75, while E or Paediatric beds are reserved for minors under 16. The alternative assumption, that age does not affect the admission rate or average LoS holds only for a small minority of cases. One of many counterexamples is the maternity ward, in which a specific age cohort is represented, while others are entirely absent.

A surrogate for national data on the age distribution for pathology-bed type combinations is a sample from a large set of Belgian hospitals. The authors derived the age distribution in sets of pathology-bed type combinations from a dataset of more than ten different hospitals. The sample of hospitals is skewed towards the Flemish region, but deemed representative. There is no clear rationale that supports the hypothesis that geographic location significantly modifies the impact of age on the admission rate or Length of Stay, and that the geographic spread in the dataset would be insufficient to neutralize it if it did. Accordingly, the age distribution and average LoS per age group derived from this large dataset is concluded to be representative and applied in the extrapolations.

The extrapolations based on historical trends pose similar challenges. National historical data on the admission rates or average LoS for pathology-bed type combinations are not available. Historical data on admission rates and LoS related to pathologies, however, are. Accordingly, the implicit assumptions in using this proxy are evaluated. The implicit assumption of deriving a historical trend on the basis of the larger pathology group is that the considered parameter of the individual bed types rises or falls proportionately with that of the related pathology groups. That is, the historical trend in admissions and average LoS, corrected for demographic differences, is the same across all bed types per pathology. Experts confirm this assumption to be reasonable.
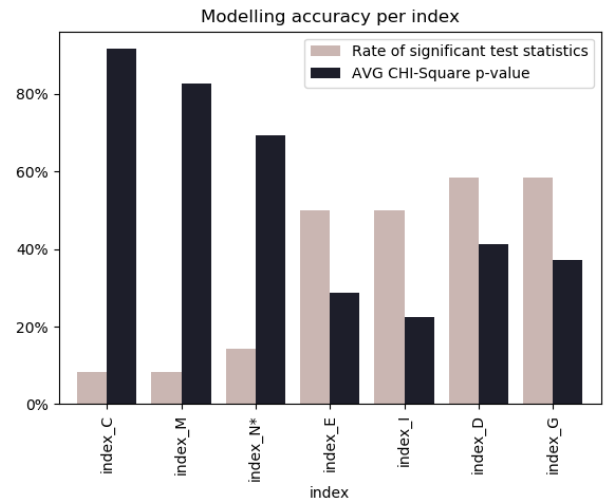

Fig. 1: Metrics of modelling accuracy per bed type

IV. Application

The methodology described in this paper is applied to a set of Belgian hospitals. First, a comparison is made between the frequency distribution of the daily inpatient bed occupancies that is observed in the data and the one resulting from the simulation in order to evaluate the validity of the methodology. Second, an aggregation is made of the different hospitals in order to review the evolution of the need for different bed types, which generally correspond with wards, in Belgium.

A. Modelling accuracy

CHI-square tests are used to compare the simulated inpatient frequency distribution with the observed distribution. In 62.3% percent of the 77 cases, the simulated distribution does not significantly differ from the observed distribution. It is noted that the rate of finding significant differences is not uniform across the different bed types, nor hospitals. Figure 1 shows that index C, M and N* yield frequency distributions that are much more often aligned than index E, I, D, and G.

There are several reasons why the simulated frequency distributions are not expected to be representative of the observed distributions. First, the bed types do not form entirely independent wards. Often, subdivisions are made according to specialty. Those subdivisions are not necessarily perfectly communicating vessels, as is assumed by the model. If they are not, it would expected that at higher occupancy levels are
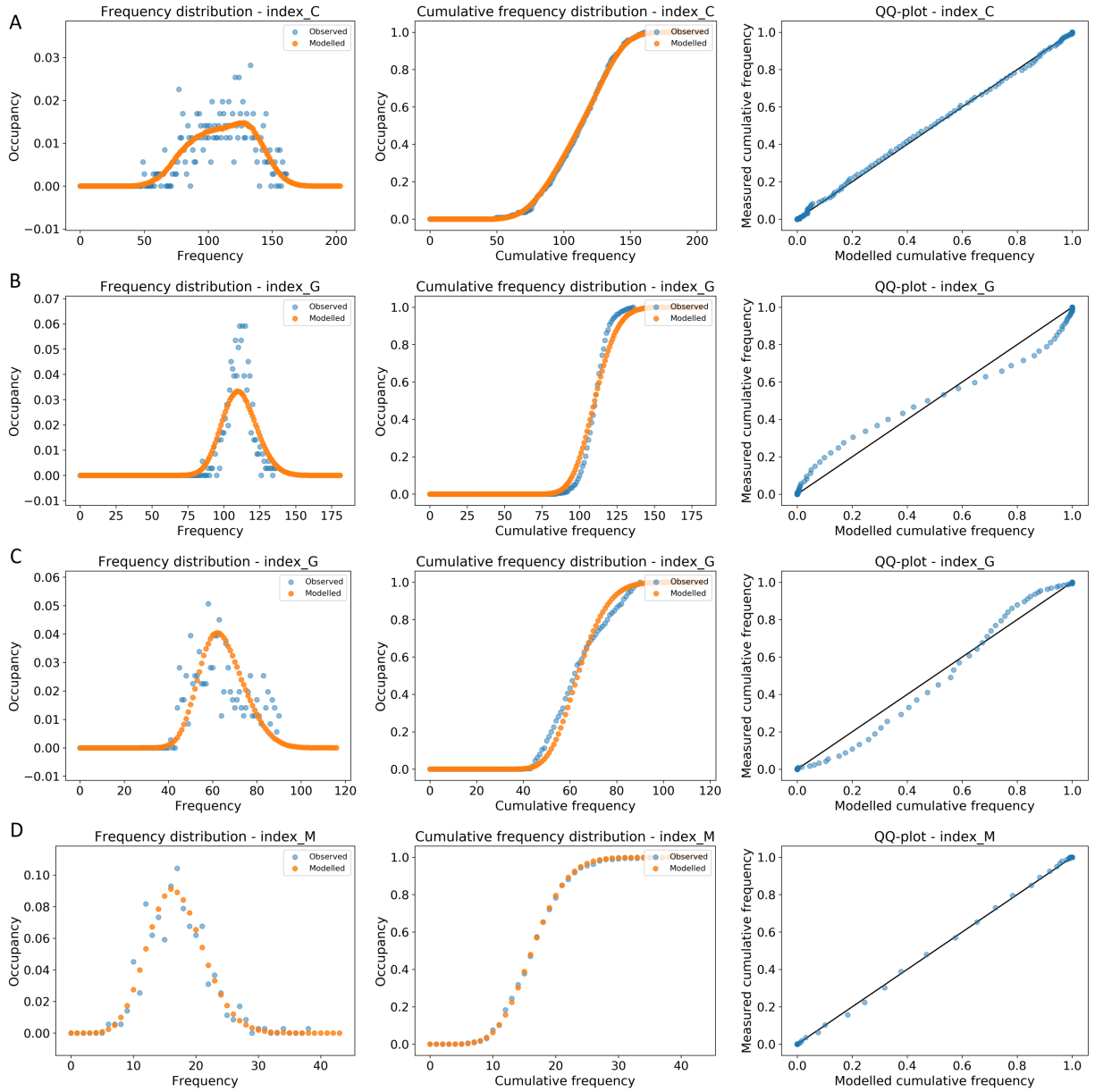

Fig. 2: Examples of distributions comparing observed and simulated pattern. From left to right: Frequency distribution of occupancy, cumulative frequency distribution of occupany, and QQ-plot comparing cumulative frequency distributions.

generally skewed down. Suppose, for instance, that a bed of type D is still available, though related to a specific specialty. Administrators might accomodate the patient elsewhere, in a bed of a different type, if the patient requires treatment from a different specialty. Thus, higher occupancies would be inhibited. Additionally, as described and modelled by Mallor et al. [6], active management of patient stays to level out capacity demand can skew patient occupancies towards the average. For instance, administrators choose to discharge patients earlier than they would on average if occupancy levels are high.

Figure 2 shows some examples of plots comparing the observed and simulated cumulative frequency distributions. For some classes, such as A and D, observed frequency distributions closely align with those empirically observed. Others, such as B, show a distribution where the occupancies are skewed towards the average. This might be caused by an overestimation of the seasonality effects, or more likely, have more systematic causes, such as active bed occupancy management. Another alternative, is that for some bed types, the planning of treatments is done in an even more coordinated way, reducing the variability of admissions as compared to the asumptions described in Latruwe et al. [5].

Another variant is C, which is rare compared to variant B, and has a wider spread than expected by the model. This might be caused by an underestimation of the seasonality effects, or due to operational conditions. It is not clear which operational conditions could lead to these results. One possibility is that some relevant seasonality effect types exist that are not covered by the month-of-year and day-of-week varieties included in the model. For instance, official holidays, and common holiday periods that do not match entire months might affect planned admissions in a way that causes the observed effect.

Additional issues with the seasonality patterns might exist. First, it is possible that seasonality patterns do not manifest themselves at the same time every year. For instance, the peak of a flu epidemic might be shifted by several weeks in one year versus another. In the current model, such discrepancies would be leveled out, flattening the expected seasonal peaks. Second,

TABLE I: Expected evolution of care demand in inpatient days by bed type, based on sample of hospitals.

	index_C	index_D	index_E	index_G	index_I	index_M	index_N*
2017	100%	100%	100%	100%	100%	100%	100%
2022	-5.6%	+0.0%	-6.7%	-1.1%	+1.4%	-9.1%	+4.7%
2032	-11.9%	+1.6%	-19.7%	+3.1%	+5.0%	-22.0%	+4.3%

the model does not insist on proof that a seasonality pattern is present in the data, rather, it assumes that there is and measures it using as much data as is available. There would be too few observations to statistically validate monthly seasonal swings except in the more extreme cases, or in the most common pathology categories. Figure 3 shows the relative confidence bounds around the mean of monthly admissions for 2016-2017, assuming admissions would follow a Poisson pattern. The expected admission rate according to the derived monthly seasonal pattern, with the day-of-week effects neutralized for clarity, are also shown. Since only two years of observations are available, the confidence intervals are large, and it is inevitable that false seasonality patterns are introduced in some cases. Including more years of observations will naturally mitigate this issue.

Thus, aside from the effects of management decisions and seasonality effects that are not included, increasing the number years included in the seasonality derivations is expected to improve the number of wards that pass the CHI-square comparative test.

B. Model forecasts

Table I shows the expected evolution of care demand in inpatient days based on the sample of hospitals available. These results can be compared to the evolution that took place in the past, and to forecasts by other authors. The direction of the past evolution of the number of beds per bed type, which does not necessarily correlate perfectly with the total amount of inpatient days per bed type, is consistent with the model's forecasts. One exception is D-beds, demand for which
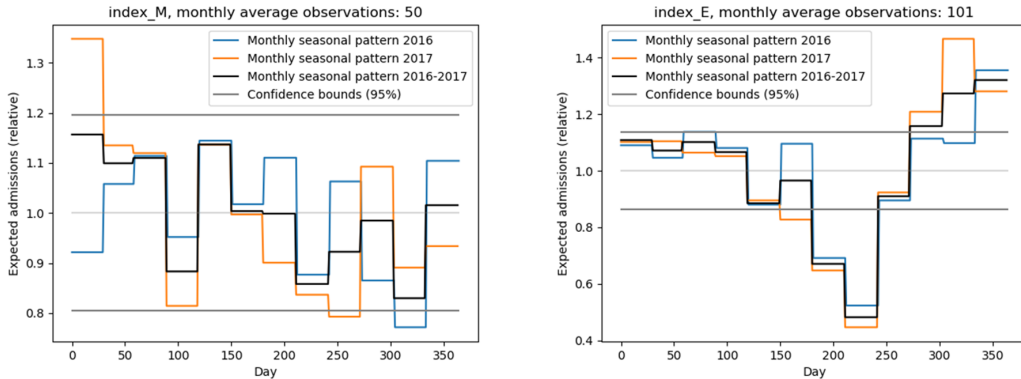

Fig. 3: Examples of seasonality effects suggested by the model, alongside confidence bounds around estimated admission numbers, in relative terms.

is expected to remain more or less stable as opposed to the decreasing trend measured in the last decade [7].

Though comparisons with Van de Voorde et al. [8] are difficult because the forecasts concern different time periods and samples, these results do not contradict their general findings.

V. CONCLUSION

This work describes a methodology to model subparts of inpatient stays per pathology for capacity need forecasts of individual hospitals and wards, while ensuring that historical extrapolation of relevant parameters and applying demographic corrections remains possible. The model is unit and facility independent. It alleviates the need to use less intricate approximation methods to tranform estimates of inpatient days per pathology, based on entire stays, to capacity estimates for wards or bed types that are only relevant for a part of the patient's stay. Additionally, it enables the derivation of inpatient occupancy frequency distributions, which can support capacity decisions using service levels as target performance indicators. Lastly, but perhaps most importantly, it provides an alternative to common hospital Discrete Event Simulation models that use transfer probabilities to model patient trajectories, and as a consequence, are forced to make abstraction of pathology details.

The similarity of simulated and observed cumulative occupancy frequency distributions, however, as measured by the CHI-square test, could further be improved. In the considered sample, the rate of non-significantly differing distributions is 62.3%. More data or further review of the seasonality mechanism could further improve the similarity metrics through the improvement of the modelling of exogenous effects, as opposed to the effects of endogenous influences, such as management interventions driven by occupancies.

ACKNOWLEDGMENT

REFERENCES

[1] P. Devapriya, C. T. Strömblad, M. D. Bailey, S. Frazier, J. Bulger, S. T. Kemberling, and K. E. Wood, "Stratbam: a discrete-event simulation model to support strategic hospital bed capacity decisions," *Journal of medical systems*, vol. 39, no. 10, p. 130, 2015.

[2] A. Komashie and A. Mousavi, "Modeling emergency departments using discrete event simulation techniques," 2005.

[3] G. W. Harrison, A. Shafer, and M. Mackay, "Modelling variability in hospital bed occupancy," *Health Care Management Science*, vol. 8, no. 4, pp. 325–334, 2005.

[4] P. R. Harper and A. Shahani, "Modelling for the planning and management of bed capacities in hospitals," *Journal of the Operational research Society*, vol. 53, no. 1, pp. 11–18, 2002.

[5] T. Latruwe, M. Van der Wee, P. Vanleenhove, J. Devriese, S. Verbrugge, and D. Colle, 2020, submitted for publication.

[6] F. Mallor and C. Azcárate, "Combining optimization with simulation to obtain credible models for intensive care units," *Annals of Operations Research*, vol. 221, no. 1, pp. 255–271, 2014.

[7] Agentschap Zorg en Gezondheid, "Cijfers: Evolutie van de ligduur en bezettingsgraad per ziekenhuisdienst - Zorg en Gezondheid." [Online]. Available: https://www.zorg-en-gezondheid.be/evolutie-van-de-ligduur-en-bezettingsgraad-per-ziekenhuisdienst

[8] C. Van de Voorde, K. Van den Heede, C. Beguin, N. Bouckaert, C. Camberlin, P. de Bekker, N. Defourny, H. De Schutter, C. Devos, S. Gerkens *et al.*, "Required hospital capacity in 2025 and criteria for rationalisation of complex cancer surgery, radiotherapy and maternity services," Belgian Health Care Knowledge Centre (KCE), Tech. Rep., 2017.