Open access • Posted Content • DOI:10.1101/2021.04.10.21255243

Hospital Length of Stay: A cross-Specialty Analysis and Beta-Geometric Model

Nassim Dehouche, Viravan S, Santawat U, Torsuwan N ...+3 more authors

Institutions: Mahidol University International College, Mahidol University

Published on: 17 Apr 2021 - medRxiv (Cold Spring Harbor Laboratory Press)

HOSPITAL LENGTH OF STAY: A CROSS-SPECIALTY ANALYSIS AND BETA-GEOMETRIC MODEL

Nassim Dehouche, Ph.D.
Business Administration Division
Mahidol University International College
Salaya, Thailand
Nassim.deh@mahidol.edu

Sorawit Viravan, M.D.
Department of Pediatrics, Faculty of Medicine Siriraj Hospital
Mahidol University
Bangkok, Thailand
Sorawit.vir@mahidol.edu

Ubolrat Santawat, M.D.
Deputy Dean, Finance, Faculty of Medicine Siriraj Hospital
Mahidol University
Bangkok, Thailand
Ubolrat.san@mahidol.ac.th

Nungruethai Torsuwan
Faculty of Medicine Siriraj Hospital
Mahidol University
Bangkok, Thailand
Nungruethai.tor@mahidol.ac.th

Sakuna Taijan
Faculty of Medicine Siriraj Hospital
Mahidol University
Bangkok, Thailand
Sakuna.tai@mahidol.edu

Atthakorn Intharakosum
Faculty of Medicine Siriraj Hospital
Mahidol University
Bangkok, Thailand
Atthakorn.int@mahidol.edu

Yongyut Sirivatanauksorn, M.D., Ph.D.
Department of Surgery, Faculty of Medicine Siriraj Hospital
Mahidol University
Bangkok, Thailand
Yongyut.sir@mahidol.ac.th

ABSTRACT

The typical hospital Length of Stay (LOS) distribution is known to be right-skewed, to considerably vary across Diagnosis Related Groups (DRG), and to contain markedly high values, in significant proportions. These very long stays are often considered outliers, and thin-tailed statistical distributions are assumed. Moreover, modeling is typically performed by Diagnosis Related Group (DRG) and is consequently based on small empirical samples, thus justifying the previous assumption. However, resource consumption and planning occur at the level of medical specialty departments covering multiple DRG, and when considered at this decision-making scale, extreme LOS values represent a significant component of the distribution of LOS (the right tail) that determines many of its statistical properties.

Through a study of $46,364$ electronic health records over four medical specialty departments (Pediatrics, Obstetrics/Gynecology, Surgery, and Rehabilitation Medicine) in the largest hospital in Thailand (*Siriraj* Hospital in Bangkok), we show that the distribution of LOS exhibits a tail behavior that is consistent with a subexponential distribution. We analyze some empirical properties of such a distribution that are of relevance to cost and resource planning, notably the concentration of resource consumption among a minority of admissions/patients, an increasing residual LOS, where the longer a patient has been admitted, the longer they would, counter-intuitively, be expected to remain admitted, and a slow convergence of the Law of Large Numbers, making empirical estimates of moments (e.g. mean, variance) unreliable. Consequently, we propose a novel Beta-Geometric model that shows a good fit with observed data and reproduces these empirical properties of LOS. Finally, we use our findings to make practical recommendations regarding the pricing and management of LOS.

Keywords: Length of Stay, Statistical Modeling, Extreme Value Theory, Beta-Geometric Distribution.

Acknowledgements

The first author is grateful to Prof N. N. Taleb for his valuable advice.

1 Introduction

In a global healthcare sector that is gradually but steadily shifting from a fee-for-service to value-based care agreements, Length of Stay (LOS) is a useful indicator of resource utilization and cost-efficiency, and has been likened to a currency for healthcare decision-making [1, 2]. Because expenses are largely fixed in the first few days following admission, the marginal value of a bed is determined by the alternate use that the healthcare provider would have made of it (i.e. its opportunity cost) [2]. This value may be neglectable for low-utilization healthcare units. However, for high-demand hospitals operating close to capacity, the opportunity cost of long LOS can be significant. This is the case for Siriraj Hospital, the oldest and largest hospital in Thailand. Siriraj being an affordable, yet very reputed healthcare provider in the country, its specialty departments typically operate at full capacity, with long waiting lists for admissions that can in some cases take several months. In such circumstances, a patient with a stay of n days virtually translates into the inability to admit n patients with an LOS of 1 day.

The average or median LOS are commonly used indicators to gauge the efficiency of a healthcare facility, and reduced LOS is associated with improved patient outcomes, decreased risks of developing healthcare-acquired infections (HAI), and a fairer access to healthcare.

As a precursor to the construction of indicators and benchmarking, understanding the statistical properties of LOS, at the relevant scale of the healthcare unit, is crucial for decision-making.

In this study, $46,364$ electronic health records over four medical specialty departments were analyzed, with a focus on the tail properties of the distribution of LOS, and their managerial consequences. The remainder of this paper is organized as follows. In Section 2, we conduct a detailed review of the related literature and classify extant studies based on their sample sizes, underlying statistical models, scope, and treatment of outliers. Section 3 describes our data and presents basic descriptive statistics for LOS in the four considered specialty departments. Section 4 defines the theoretical framework of this paper, which consists in various methods from the Extreme Value Theory toolbox and the Beta-Geometric model we propose for LOS. Section 5 presents our results and Section 7 further elucidates their implications for profit margins and resource management. Finally, Section 8 concludes this paper with general remarks concerning our statistical analysis and its limitations.

2 Related Work

The typical Length of Stay (LOS) distribution is known to be right-skewed, to considerably vary across Diagnosis Related Groups (DRG) [7], and to contain markedly high values, in significant proportions. These characteristics make the use of thin-tailed models and least-squares inference methods based on the Gauss-Markov theorem [4] hard to justify. Moreover, they make the calculation of averages less reliable and representative of the typical observation [5]. Extant works on the statistical modelling of LOS attempt to circumvent these difficulties, with the following solutions, which are reviewed in this section:

1. Using different models of LOS for different DRG in a care unit.
2. Trimming or discarding outliers.
3. Using different models for short and long stays (i.e. mixtures of distributions).
4. Relying on heavy-tailed models.

Moreover, the typical statistical study of LOS either uses relatively small empirical datasets (often as a result of point 1.), which may artifically make point 2. appear justified, or simulated data with embedded assumptions regarding the distribution. Lastly, it seldom distinguishes measures of LOS per patient, from LOS per admission. This point could have its importance in analyzing the effect of readmissions (i.e. admissions within 30 days of discharge) or multiple admissions (admission after more than 30 days of discharge) on resource consumption.
Table 1 summarizes the characteristics of the main references for statistical models of LOS in the literature.

2.1 Scale of modeling

According to Ickowicz et al. [14], *Diagnosis Related Groups* (DRG) have been partly created to have a form of homogeneity in resource consumption of services and costs closely related to LOS. Models of LOS at the DRG level can

Reference	Dataset size (N)	Model	By DRG	Outliers
[10]	$3,472$	Gaussian	No	Separated
[5]	$4,758,347$ (in 5 countries)	Log-normal, Weibull, Gamma	Yes (478 DRG)	No
[7]	560	Mixture of Gaussian	Yes	Trimmed
[14]	Simulated	Mixture of continuous and discrete	Yes	No
[15]	340	Log-Normal	No (ICU)	No
[16]	Unknown (in 1 hospital)	Mixture of thin and heavy-tailed	Yes (5 DRG)	No
[17]	$137 + 469$	Mixture of Exponential and heavy-tailed	Yes (1 DRG)	No
[18]	$53,965$	Gaussian	No	Trimmed
[19]	$101,766$ (in 136 hospitals)	Heavy-tailed compounds	No (diabetes)	No
[2]	1901 (in 2 hospitals)	Markov chain	No	No
This paper	$46,364$ (in 1 hospital)	Beta-Geometric	No	No

Table 1: Characteristics of extant statistical models of LOS

be valuable for understanding the specific determinants of LOS, seen as a proxy for the severity of a patient's condition, and are often used with a predictive intent. Association with independent variables of a demographic, lifestyle, or medical nature is often favored and is used to predict an individual's expected LOS.

However, this approach results in smaller samples by design, which can make very high LOS values indeed appear as neglectable outliers. Moreover, there exist more than 467 different DRG, and combining even excellent LOS models for different DRG is non-trivial and does not necessarily yield valid models for resource management at the level of a care unit covering multiple DRG, due to the complex non-linear nature of healthcare systems [20]. Thus, this approach transfers the modeling difficulty towards re-aggregating results (e.g. averages or medians per DRG, linear regression forecasts) at a scale that can inform decision-making. A task as simple as estimating the variance, coefficient of variationn, and kurtosis of LOS at the level of a care unit, based on the distribution of LOS by DRG is a non-trivial problem, even when all distributions are known to be Gaussian [6], which is far from being the case for LOS. Besides, the non-transitivity of (Pearson's) correlation [21] makes associative inference on the parts not necessarily scalable to the whole.

Ickowicz et al. [14] note that the "heterogeneity of LOS poses a problem for statistical analysis, limiting the use of inference techniques based on normality assumptions since a large number of DRGs must be analyzed routinely, automatic procedures are needed for conveniently treating skewness" and add that "the main issue is that the assumption of heterogeneous sub-populations would be more appropriate than single DRG populations". Their proposed solution is the use a mixture of probability distributions. The same point has been made by Atienza et al. [16], who conclude that "the assumption of heterogeneous sub-populations would be more appropriate than single DRG populations".

2.2 Treatment of outliers

Grubbs [23] defines *outliers* as observations that "appears to deviate markedly from other members of the sample in which it occurs".

For the positive random variable that is LOS, outliers are conventionally considered to be admissions or patients with a remarkably high LOS. Under Gaussian assumptions, empirical Length of Stay (LOS) datasets are commonly described as containing outliers.

Perhaps as a side-effect of the small samples sizes typically considered in the above studies, the notion that high-LOS admissions are outliers appears founded. Indeed, in their application using empirical data, [7] consider $N = 560$ observations clustered within 21 hospitals, with the numbers of patients ranging from 3 to 196 per hospital.

However, what are considered outliers for individual DRG may be statistically very significant in the study of LOS as a whole. Trimming/discarding outliers from the parts may discard important statistical information about the tail properties of LOS. Lee et al [7] transparently note that the use of trimming methods for outliers limits the usefulness of the models, they add "if the goal of the analysis is plan or hospital comparison, uniform trimming of LOS across all hospitals might be inappropriate in the contexts of quality improvement and performance assessment".

Ad et al. [10] is typical of such a modeling approach, which the authors remark in [11] is not of their own design but a more than 30 years old standard risk model of the Society of Thoracic Surgeons [12]. High LOS admissions are known to have a prominent financial impact. This point is stated by [31], which proposes a method that notably serves at "detecting outliers".

Lee et al. [7] recommend relying on median rather than mean estimates in the analysis of LOS, because the latter approach is more "robust to high-LOS outliers". However, the robustness of the method is, circularly, gauged on Gaussian simulated data with Bernoulli noise.

Trimming outliers, that is separating data into normal and high-LOS is commonly performed. Thus, the question of how to define thresholds for what is considered long or short LOS is important from an operational and financial point of view and at the center of many studies. In [9], the threshold for what is considered long LOS is set at 7 days or as the top 2% of LOS values which has been qualified as "somewhat arbitrary but has been applied by others in the analysis of administrative data" [7, 8]. "The objective of trimming coupled with transformation is to minimize the effects of extreme outliers and to attain the normality assumption on the LOS distribution" [7], as well as "to minimize the effects of extreme outliers and to avoid analytical problems" [9]. Leung et al. suggest trimming LOS observations at the threshold of 7 days.

The $68\% - 95\% - 99.7\%$ rule is also commonly used [5], and outliers are defined as observations that lay at a distance of more than three standard deviations from the mean, thus assuming a Gaussian distribution of LOS.

Two important points of the present article are that the notion of outliers does not apply to the type of heavy-tailed distribution that we suggest LOS belongs to, and that these so-called outliers or tail observations contain valuable statistical information. The heavier the tail of a distribution, the more statistical information it contains relative to the body [57]. Using large datasets covering four medical specialties, we notably show, that the observations considered outliers in thin-tailed models are in fact too numerous to be that, and that they have a significant impact on resource consumption and revenue.

2.3 Use of mixture distributions

A *mixture distribution* is the probability distribution of a random variable that is sampled from two or more different probability distribution functions (PDF) [6]. The PDF of the mixture random variable is often a weighted sum of the PDFs of the mixed random variables. We distinguish *mixture distributions* from *compound distributions*, the latter being distributions whose parameters are themselves random variables [30].

The problem with mixture probabilities may be overfitting, especially with the small sample sizes that are typically considered. As noted in [11], this is an important pitfall in modeling LOS.

Atienza et al. [16] divide patients by DRG and use a mixture of thin and heavy-tailed distributions (Gamma, Weibull, Log-Normal). Simulated data of 128 samples of size 100 ($N = 12800$) and 100 samples of size 500 ($N = 50000$).

Gardiner et al. [17] consider a first sample of $N = 137$ patients who underwent bone marrow surgery and a second sample of $N = 469$ patients in Psychiatry. In the latter sample, they interestingly report LOS varying from 1 to $24,028$ days, with a mean of $3,712.4$ days and a median of $1,134$ days. The authors recommend a mixture of exponential and heavy-tailed distributions (Pareto, Generalized Pareto).

Lee et al. [7] consider one DRG in Obstetrics and Gynaecology ("Cesarian delivery with severe complicating diagnosis"). They consider an empirical dataset of $N = 560$ patients. Outliers are trimmed. They use a mixture of (Gaussian) distributions.

Ickowicz et al. [14] consider a model for "short stays" and a different model for "long stays". Moreover, a mixture of continuous (Normal or Log-Normal) and discrete (Poisson, Binomial, Negative Binomial) random distributions. Their sample size is unknown but it is mentioned that is similar to that of [16], which consists in simulated data.

Rady et al. [15] consider $N = 340$ patients of an Intensive Care Unit. No division by DRG is performed, however the minimum observed LOS was 1 day, and the maximum was 60 days.

2.4 Treatment of tails

A distribution is said to be *heavy-tailed* if its survival function $S(t)$ satisfies $e^{\lambda \cdot t} S(t) \to +\infty$ as $t \to +\infty$ [17]. Its moments, including the mean, may or may not be finite. A distribution is called thin or light-tailed otherwise.

Gardiner et al. [17] make some important remarks about outliers and tails. They note that "in modeling hospital Length of Stay (LOS) and inpatient cost, extreme values in the data are likely and should not be regarded as outliers for deletion or downweighted in analyses".

Hospital Length of Stay (LOS) is often modelled with thin-tailed distribution for the purpose of least square regression.

A recent debate [11] concerning the statistical representation of LOS is a thoracic surgery department [10] highlighted the prevalence of thin-tailed statistical modeled in this context.

Though no assumptions are explicitly made regarding the distribution of LOS, [10] use linear regression on averages and medians which point to Gaussian underlying distribution. The former point was criticized by [11] which observes that "ordinary least squares regression does not adequately accommodate large LOS values" and recommends "normality-improving data transformations", which typically consist in Log-transformations. However, [2] point out that the weakness of this approach is that "Log-LoS" does not have an intuitive meaning and is therefore not useful for policy making. Moreover, the retransformation of the regression results from log results through exponentiation is complicated by heteroscedasticity (i.e. the fact that the variability of LOS is unequal across the range of values of the independent variables used to predict it), and produces "very imprecise estimates if the log-scale error is heavy-tailed". In fact, single-point forecasts are not theoretically justified for heavy-tailed variables with high standard deviation [3]. Moreover, the mean and other moments may be infinite or require a very large number of observations to be estimated, due to the slow convergence of the Law of Large Numbers for variable of this type. This point has also been highlighted by [17].

A more robust approach consists in modeling the relevant properties for decision-making (e.g. through fitted distributions) rather than predicting punctual values.

Marazzi et al. [5] assessed the adequacy of three conventional parametric models, Log-normal, Weibull and Gamma, for describing the LOS distribution. But, as Lee et al. [7] point out, none of them seemed to fit satisfactorily in a wide variety of samples. Moreover, [2] finds that

Harini [19] use data from a public repository covering 136 hospitals $N = 101,766$. All diabetes patients. No division by DRG is performed beyond that. They recommend heavy-tailed distributions (Beta-Cauchy, Gamma-Pareto, Gamma-Exponential-Cauchy). It should be noted that these are compound distributions, not mixtures of distributions. Moreover [5] suggest two approaches to respond to the skewed nature of the distribution of LOS and building "outliers resistant (robust) methods": Using other transformations than the Log-normal or using other types of distribution. The authors pursue the latter approach and test three heavy tailed distributions (Log-normal, Weibull, Gamma), which are found to adequately fit the distribution of LOS, over 3279 samples totaling approximately 5 million stays in multiple European countries. Among these, the Log-normal model is found to fit the majority of samples.

Based on a dataset recording 1901 patients' LOS, Faddy et al. [2] develop an intuitive continuous Markov process which is found to provide a better fit than Gamma and Log-normal models. This model is based on the assumption that each day of admission, patients either progresses to another day of hospital stay, corresponding to increasing their LoS, or they are discharged (absorbing state). This defines a general class of probability distributions describing the random time that elapses before the absorbing state is reached. The Beta-Geometric model of LOS developed in the present article is based on the same intuition. Our model rests on the two assumptions that once admitted a patient has a certain likelihood of being discharge any subsequent day, and that different patients have different such likelihoods.

However, because of the discrete nature of LOS, we consider a discrete Beta-Geometric process. Thus, our model differs and complements the reviewed literature as follows:

- We consider large samples of empirical observations ($N = 46,364$, over four medical specialty departments).
- The Beta-Geometric is discrete, right skewed and can fit thin to very heavy tail behavior, depending on the choice of parameters of the distribution.
- The object of our analysis are both the LOS per admission as well as the LOS per patient, highlighting the effect of multiple admissions/readmissions on the statistical properties of LOS.
- The scale of our model are care units corresponding to four medical specialties, but the model can be scaled up (e.g. to a hospital) or down (e.g. DRG) without loss of generality.
- We highlight and model important properties for resource management, such as the behavior of the mean excess function, Maximum to Sum ratios and the effect of the concentration of LOS among a minority of admissions/patients on bed turnover and revenue.
- We propose various novel graphical tools to complement LOS analysis. Notably the use of Mean Excess Plots to detect mixtures of probabilities in and thresholds.

3 Data Description and Basic Statistical Properties

The dataset used in this study covers four medical specialty departments at Siriraj Hospital. These are Surgery, Obstetrics and Gynaecology (OB/GYN), Pediatrics, and Rehabilitation Medicine (Reh. Med.). The present study was approved (Certificate of Approval No. Si 032/2021) by the Institutional Review Board of Mahidol University. All data provided to the first author for this study were de-identified.

For the purpose of this study, we utilize the following data fields:

- Admission Number: A unique identifier for each admission.
- Patient ID: A unique identifier for each patient.
- Admission Date.
- Discharge Date.
- Length of Stay: The difference in days between Discharge Date and Admission Date, the minimum value of which is 1 days. Consequently, random variables modeling LOS are truncated if having support on negative values, and shifted by one day, if having support on zero.
- Total Charge: The total billing expense to the payer, in Thai Baht (THB), resulting from an admission.
- Discharge Status: We distinguish positive discharge outcomes ("Complete recovery", "Improved", "Delivered", "D/C with mother") and negative ones ("Dead", "Not improved", "D/C separately"), when applicable to admissions in a medical specialty.

Throughout this paper, we distinguish between LOS per admission (number of days between admission and discharge for one admission, identified by an admission number) and LOS per patient (cumulated LOS for all admissions of a patient identified by a patient ID). These two levels of analysis notably allow us to study the effect of readmission and multiple admissions on the tails of the distributions of LOS. Realizations of both of these integer random variables, along with complete administrative and medical data are recorded for discharges that occurred between $13/11/2017$ and $30/09/2018$, the earliest recorded admission having occurred on $15/02/2017$ and the latest on $29/09/2018$.

Table 2 presents the sample sizes as well as basic descriptive statistics in each medical specialty department. We note that all distributions fail the Skewness-Kurtosis test for Truncated Gaussians [22]. Moreover, we note a markedly higher mean and median LOS per admission and per patient in Rehabilitation Medicine compared to the three other specialties, a fact that was also observed in [18]. These two points are related since the time-frame of data collection is the same for all specialty departments. Additionally, Rehabilitation Medicine department has a smaller capacity. Further, this department shows the smallest ratio of patients-to-admissions, indicating frequent multiple admissions over the time-horizon of the study (10.5 months). Indeed its 994 total admissions are attributed to 304 patients (patients-to-admissions ratio of 30.52%), to be compared with the ratios of 87.75% in OB/GYN, 82.92% in Surgery, and 67.82% in Pediatrics.

Moreover, the distribution of LOS in this medical specialty exhibits a significantly lower Kurtosis. Indeed, Kurtosis in the distribution of LOS in Surgery, OB/GYN and Pediatrics is extremely high and constitutes a first, important signal of its heavy-tailedness, which allows us to safely reject thin-tailed distributions (Gaussian, Poisson, etc.) at this point for these three departments. Among these three specialties, OB/GYN markedly differs by its lower Index of dispersion (i.e. variance to mean ratio). Moreover, Kurtosis significantly increases when aggregating LOS per patient, except for this medical specialty. Lastly, we note the remarkable stability of the Median LOS in Surgery, OB/GYN and Pediatrics, as a likely result of the adherence to Benchmarks in the management of LOS.

Statistic	Surgery	OB/GYN	Pediatrics	Reh. Med.
Nº of Admissions	17949	19922	7499	994
Mean	6.5973	3.8533	8.2515	24.5824
Median	4	3	4	21
Variance	125.6517	11.0901	264.5326	187.2465
Dispersion	19.0459	2.8780	32.0587	7.6170
Kurtosis	139.0954	183.9783	47.8032	4.7843
Skewness	8.9926	9.4236	5.9014	1.9304
Max	304	119	272	87
Nº of Patients	14884	17483	5086	304
Mean	7.9558	4.3908	12.1663	80.3782
Median	4	4	4	66
Variance	215.4532	15.9265	715.3997	2613.087
Dispersion	27.0810	3.6271	58.8015	32.5098
Kurtosis	279.3174	134.4066	56.51553	6.428852
Skewness	12.18991	8.437034	6.163459	2.278371
Max	546	119	451	321

Table 2: Descriptive statistics for the LOS per admission and per patient in the four departments

Cumulative Distribution Functions of LOS per admission further confirm the singular nature of Rehabilitation Medicine relative to the three other specialty departments. Figures 1, 2, 3 present the Cumulative Distribution Functions (CDF)

for LOS per admission in Surgery, OB/GYN, and Pediatrics and additionally indicate the heavy-tailedness of these distribution by comparison with typical thin-tailed CDF. However, the sigmoid CDF of LOS in Rehabilitation Medicine is not inconsistent with a thin-tailed distribution. It resembles the CDF of a Gaussian, Poisson, or Geometric distribution, but could also be that of a Beta-Geometric distribution exhibiting low variance in its underlying Beta distribution.

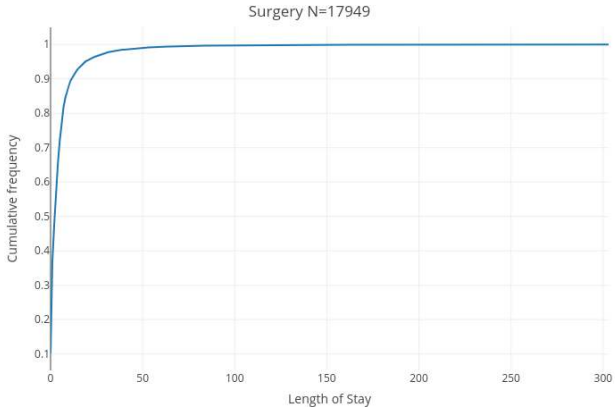

Figure 1: CDF of LOS per admission in Surgery

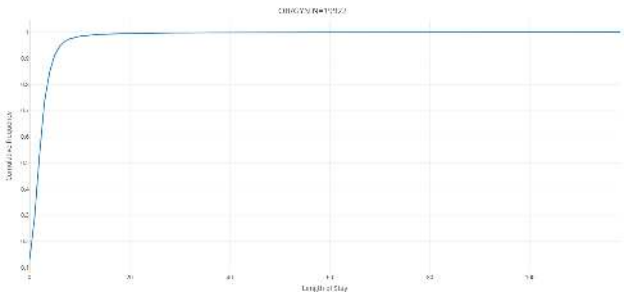

Figure 2: CDF of LOS per admission in Obstetrics and Gynaecology

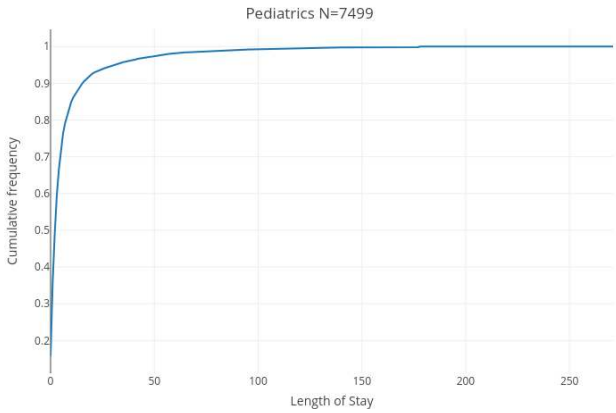

Figure 3: CDF of LOS per admission in Pediatrics

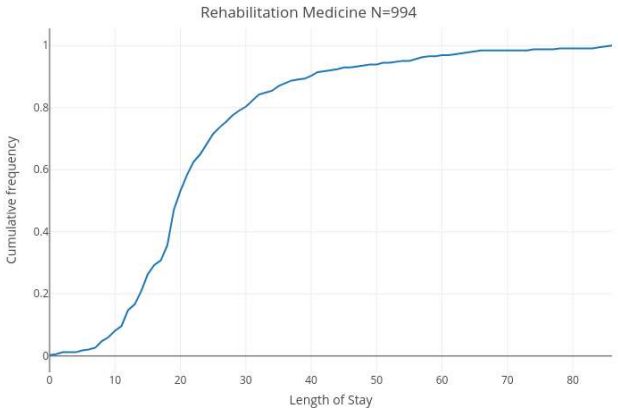

Figure 4: CDF of LOS per admission in Rehabilitation Medicine

Figure 8 presents the histogram of LOS, per admission, in Rehabilitation Medicine over which the maximum likelihood truncated Gaussian fit is superimposed (blue curves). We observe a strikingly high frequency for the particular LOS value of 20 days, which for once fits the definition of an outlier as "an observation that differs significantly from other observations" [23]. Indeed, this value shows a frequency of 144 admissions, which represents 11.46% of the 994 admissions in Rehabilitation Medicine. This markedly differs from neighboring values of 19 days with a frequency of 48 admissions (4.8% of all admissions), and 21 days with a frequency of 60 admissions (6.0% of all admissions). A LOS of 20 days was previously found by Chatterjee et al. [59] to appear in remarkably similar relative proportions to 19 and 21 days (about twice as frequently) at the Skilled Nursing Facilities (SNF), where $220,037$ discharges were made on day 20, compared to $131,558$ and $121,339$, on days 19 and 21, respectively. The explanation for this common anomaly is that the first 20 days of rehabilitation are covered in full, and patients start paying out-of-pocket from day 21, which motivates discharge on the 20th day of care based on a concern for a patient's ability to pay rather than than their recovery status. Additionally, this arbitrary but "round" value may more simply represent a psychological anchor and its frequency may be explained in an analogous manner to Benford's law for digits [60]. In any case and from a statistical perspective, the fact that some stays are artificially cut short at 20 days suggests that LOS would exhibit heavier tails without this artificial restriction. Figure 12, which represents the LOS per patient in the same specialty department shows a similar effect, with a cumulated LOS of 60 days being the overwhelmingly most frequent value (a frequency of 32 observations or 10.52% of the 304 patients) compared to the two neighboring values of 57 and 63 cumulated days (both observed 17 times or 5.59% of the 304 patients).

Though the distribution of LOS in Rehabilitation Medicine exhibits the thinnest right tail of all specialty departments, we can observe in Figure 8, and even more so in Figure 12, that it does not fit a Gaussian distribution. Indeed extreme values (e.g. above 60 days for LOS per admission, and above 200 cumulated days for LOS per patient), which would be close to impossible to observe in a Gaussian, Poisson, or Geometric distribution of similar mean of variance, are empirically way too frequent. The inadequacy of thin-tailed models is, as expected from Kurtosis, exceedingly more pronounced for Surgery, OB/GYN, and Pediatrics as can be seen in Figures 5, 6, and 7 representing the histograms of the respective distributions of LOS per admission in these departments, as well as Figures 9, 10, and 11, representing the histograms of their respective LOS per patient, over which the best fitting truncated Gaussian curves are superimposed.

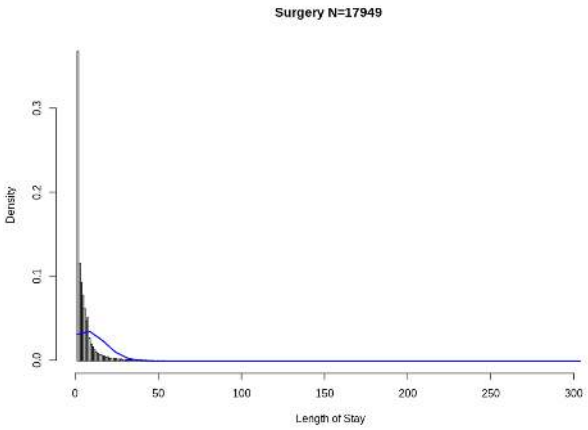

Figure 5: Histogram of LOS per admission in Surgery

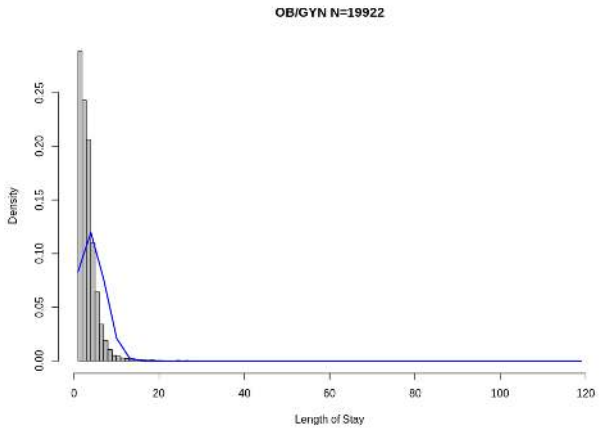

Figure 6: Histogram of LOS per admission in Obstetrics and Gynaecology

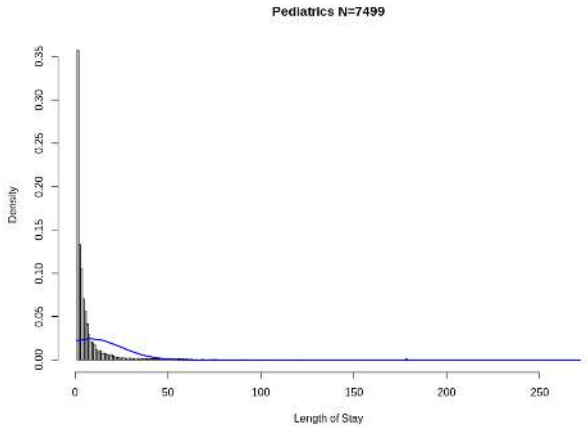

Figure 7: Histogram of LOS per admission in Pediatrics

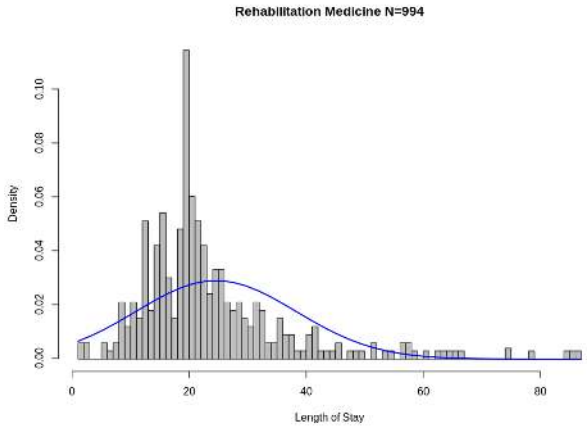

Figure 8: Histogram of LOS per admission in Rehabilitation Medicine

4 Methods

4.1 Gini Index and Lorenz Curves

The Gini index and Lorenz curves are classical statistical indicators of inequality. The Lorenz curve [36] is a representation of the CDF of a random variable showing the proportion of its total value that is concentrated in the bottom $x\%$ of observations. It is often used to represent income distribution, where it shows the percentage $y\%$ of the

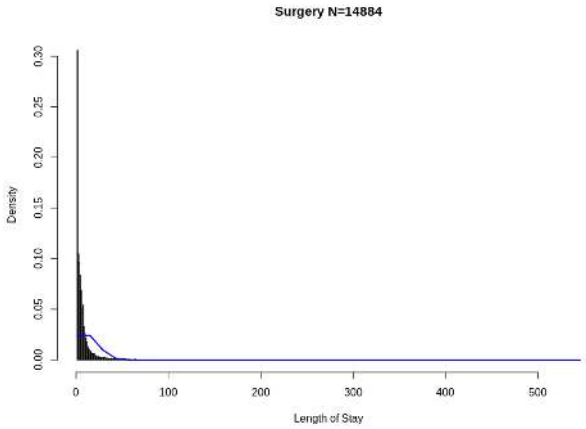

Figure 9: Histogram of LOS per patient in Surgery

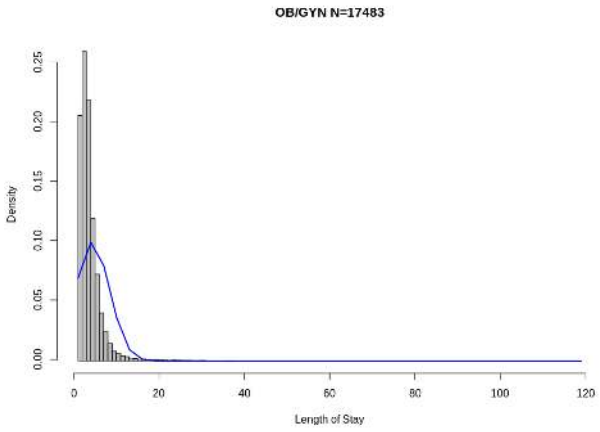

Figure 10: Histogram of LOS per patient in Obstetrics and Gynaecology

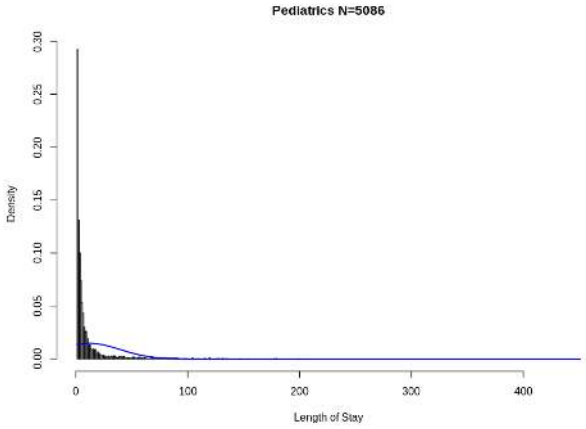

Figure 11: Histogram of LOS per patient in in Pediatrics

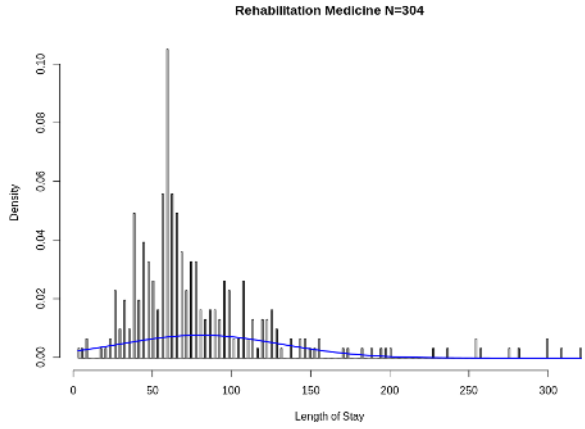

Figure 12: Histogram of LOS per patient in Rehabilitation Medicine

total wealth owned by the bottom $x\%$ of households. Inequality is represented as the area separating the Lorenz curve from the $y = x$ line, corresponding to the Lorenz curve of the Dirac delta distribution, i.e. perfect equality.

The value of the Gini index [36] represents the percentage of the area between the line of perfect equality of distribution and the observed Lorenz curve. Its possible values range from 0 to 1, higher values indicating less equal distributions. The "80-20" Pareto Principle, for instance, is indicated by a value for the Gini index of approximately 0.76 [37].

4.2 Maximum to Sum Ratios

Heavy-tailed distributions raise the question of the convergence of empirical moments (the first and second moments being the mean and variance) [42]. Given an order $p \in \{1, 2, 3, 4, \dots\}$, the convergence of the ratio of the maximum to the sum of exponent p is indicative of the existence of the moment of order p, and if so of the speed of convergence of the empirical moments of order p to its true value. Formally, given a sample of n observations $\{x_1, \dots, x_n\}$ of a positive random variable X, let $M(n, p) = Max\{x_1^p, \dots, x_n^p\}$ be the maximum of order p and $S(n, p) = \sum_{i=1}^{n} x_i^p$, the sum of order p. We have the following result [42, 43]:

$$E(X^p) < +\infty \Leftrightarrow \lim_{n \to +\infty} \frac{M(n,p)}{S(n,p)} = 0$$

Based on the previous equivalence, Maximum to Sum plots [45, 44] represent the ratio of the maximum to sum of order p as a function of the number of data points for different values of p and indicate a convergence of the moments of order p to a finite value if and only if the ratio converges to zero. Figures 13, 14, 15, and 16 present the classical behavior of such plots for Monte-Carlo samples of 10^5 observations from the thinest-tailed to the heaviest-tailed distributions. These figures respectively illustrate thin-tailed, memoryless, sub-exponential, and fat-tailed classes of distributions.

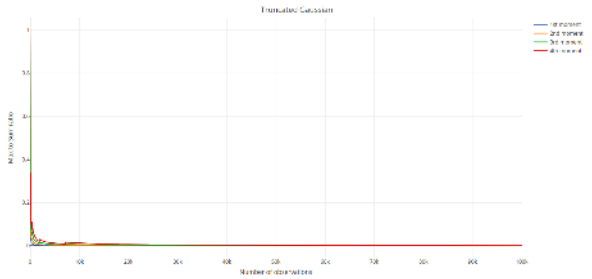

Figure 13: Maximum to Sum ratios for a Gaussian

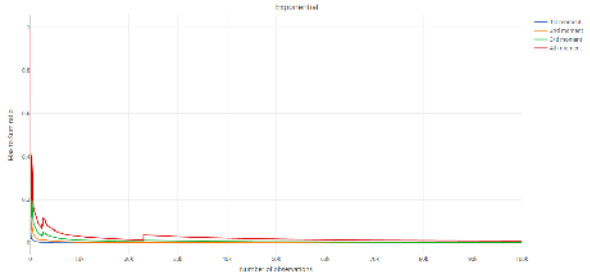

Figure 14: Maximum to Sum ratios for an Exponential

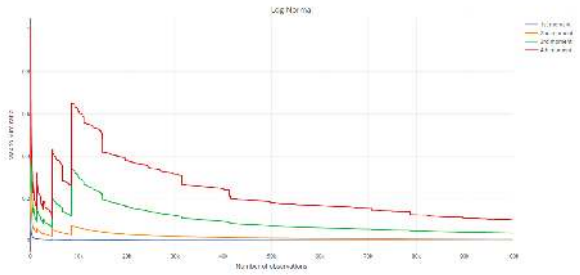

Figure 15: Maximum to Sum ratios for a Beta-Geometric with $\alpha = 1.3$ and $\beta = 3$

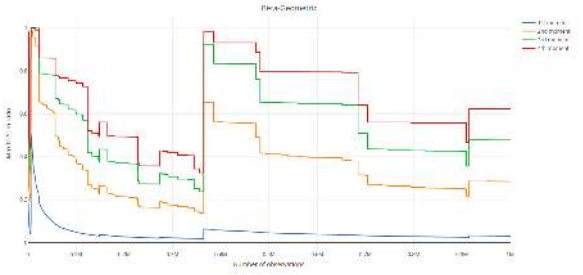

Figure 16: Maximum to Sum ratios for a Beta-Geometric with $\alpha = 1$ and $\beta = 3$

4.3 Mean Excess Functions

The excess distribution over a threshold a for a random variable (such as a duration, e.g. LOS), with support in $D(X)$, is defined [52, 51] as $F_a(x) = P(X - a \leq x | X > a), a \in D(X)$. Intuitively, its complement $1 - F_a(x)$ measures the likelihood of X exceeding $a + x$, given that X has exceeded a. For instance, if X measures LOS, $1 - F_a(x)$ is the likelihood of a patient staying x more days, given that they have been admitted for a days so far. The Mean Excess function, also known as the Mean Residual Life function, is the expectation of this distribution for random variable of finite expectations and is defined as $ME(a) = E(X - a | X > a), a \in D(X)$. If X measures LOS, this would be the expected remaining LOS of a patient, given that they have been admitted for a days so far.

This excess distribution and mean are the foundations for peaks over threshold (POT) modeling [52] which fits distributions to data on excesses and has wide applications notably in risk management, actuarial science, project management, and survival analysis. Moreover, they define three classes of random variables whose life-expectancy exhibits crucially different statistical behaviors:

- A decreasing mean excess function is characteristic of thin-tailed random variables with memory. If the variable measures the duration of a certain state, the longer an object has been in that state, the lower the expected remaining duration. For instance, the number of miles driven until the engine of a car breaks down is in this class. Gaussian or Poisson random variables possess this property.

- A constant mean excess function is characteristic of *memorylessness* [54]. Exponential random variables and their discrete analogues, Geometric random variables, notoriously exhibit this property. For instance, if a store clerk hasn't seen any new customers after 30 minutes of opening, the conditional probability that a customer showing up will take at least 10 more minutes is equal to the unconditional probability of initially waiting 10 minutes after opening.
- An increasing mean excess function is characteristic of scalable heavy-tailed random variables and corresponds to the *Lindy Effect* [26] where "the longer you wait, the longer you will be expected to wait". For instance, the longer a book has been in print or a project has been running late, the longer their expected remaining duration in a state of print or tardiness respectively, and any additional period in those states increases the expected remaining duration. Lindy things "age in reverse" [58].

Figures 17 and 18 illustrate the distinction between these three classes with Monte Carlo simulated samples of 10^6 observations from each distribution. In Figure 17, we plot the Mean Excess function of a truncated Gaussian (blue line) of mean 10 and variance 3 and Exponential (red line) of rate .1. Note that the chaotic perturbations at the extremity of the red curve are just the effect of the finite sample bias, i.e. the fact that points for very high order statistics in the plot are the result of very few observations [40].

Figure 18 present the mean excess functions of the exponential of the previous Gaussian, i.e. a Log-Normal distribution (green line), as well as a Pareto distribution of shape parameter .2.

The axis labels have been voluntarily omitted to fit both curves within one plot and highlight the monotonic shape of the functions rather than specific values.

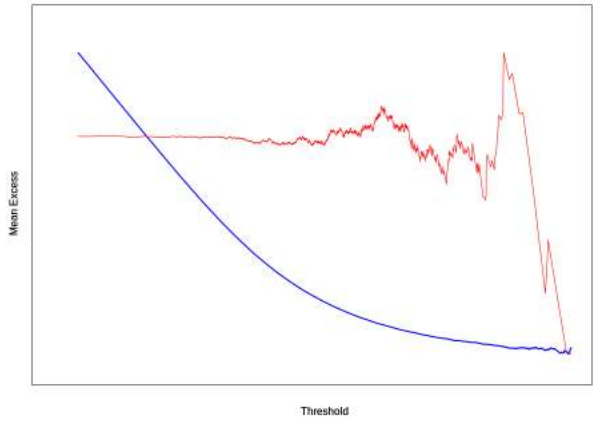

Figure 17: Mean Excess Functions for a Gaussian and Exponential

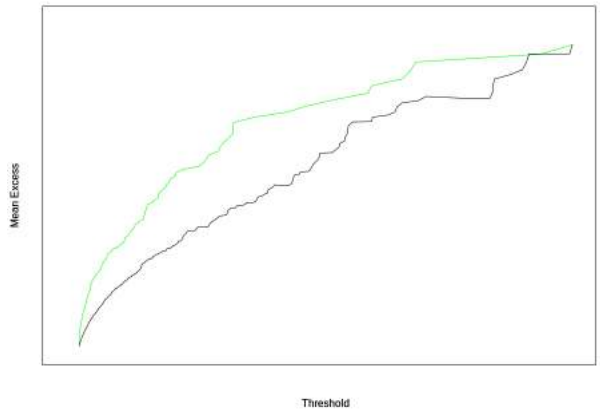

Figure 18: Mean Excess Functions for a Pareto and Sub-Exponential

4.4 The Subexponential Class of Distributions

Let $X = X_1, \ldots, X_n$ be a sequence of positive independent and identically distributed random variables with cumulative distribution function (CDF) F. Following [24, 25, 41], we consider that X belongs to the subexponential class if it satisfies the following property:

$$\lim_{x \to +\infty} \frac{1 - F^{(2)}(x)}{1 - F(x)} = 2$$

Where $F^{(2)}$ is the CDF of $X_i + X_j, i, j \in \{1, \ldots, n\}$, the sum of two independent copies of X. Practically, this property implies that likelihood of two independent observations of X (e.g. two patients' LOS) exceeding a high threshold x is twice the likelihood of either one of them exceeding x. Thus, for only two observations, the value of the sum, if high, is dominated by an individual observation and the other one contributes negligibly. This property can be extended to n observations [24, 25, 41], where the property $\lim_{x \to +\infty} \frac{1-F^{(n)}(x)}{1-F(x)} = n$ also characterizes the subexponential class. Therefore the sum of n observations would be dominated by extreme values, which makes

aggregate indicators based on the sum (e.g. the mean) less indicative of a typical value and more sensitive to extreme values. In distributions verifying this property, extreme observations can disproportionately impact and determines sums or aggregates, an example of which being the wealth of a group of people in which the wealthiest person in the country is included [26]. The total or average wealth of the group would be overwhelmingly determined by the wealth of that person. This property is also known as the *catastrophe principle*[41]. For subexponential distributions, the simplest explanation for a large sum and thus mean is that one large observation happened, not that a collection of many slightly larger than expected events conspired together to make the sum large. This property runs completely contrary to what happens under model thin-tailed distributions considered to model length of stay (Gaussian, Poisson), but is present in Log-normal models and the Beta-Geometric model we propose herein.

Another important consequence of this property is the inapplicability of the Gauss-Markov theorem [24] and thus of linear least-squares regression methods. However, maximum likelihood estimation methods [34] are applicable in this context [24].

4.5 The (shifted) Beta-Geometric Distribution

Our proposed model for hospital Length of Stay is based on the following two assumptions, in which X is a random variable representing its value for an individual admission or patient:

1. Once admitted, a patient has a constant probability p of being discharged any subsequent day. Given p, this is equivalent to assuming that X follows a (shifted, i.e. starting from 1 instead of 0) Geometric distribution, with probability density function (PDF) $P(X = x|p) = p \cdot (1-p)^{x-1},\ x \in \{1, 2, 3, \dots\}$
2. Patients within a medical unit, department, or hospital exhibit different probabilities of discharge p depending for instance on their diagnosis related group, individual health condition, type of care, hospital policy, and other factors. In other words, p itself is a random variable. The Beta distribution [29] is conventionally used to model the variations of a probability in a population. It is characterized by two positive real parameters α and β and gives the following PDF for p: $f(p|\alpha, \beta) = \frac{p^{\alpha-1} \cdot (1-p)^{\beta-1}}{B(\alpha,\beta)},\ p \in [0, 1]$, where $B()$ is the Beta function given by $B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}$ and $\Gamma()$ is the Gamma function given by $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} \cdot e^{-x} dx$.

Assuming a count that starts at 1 and under the two above assumptions X follows a (shifted) Beta-Geometric distribution, a compound distribution characterized by the two parameters $\alpha, \beta \in \mathbb{R}^+$, with the following PDF:

$$P(X = x|p) = \frac{B(\alpha + 1, \beta + x - 1)}{B(\alpha, \beta)},\ x \in \{1, 2, 3, \dots\}$$

It should be noted that the Beta-Geometric distribution is not a mixture of two probability distributions (Beta and Geometric), but a coherent compound distribution that results from a geometric process with variable success probability.

Despite its simplicity, the Beta-Geometric distribution is very versatile. This distribution can exhibit a wide variety of tail behaviors, depending on the parameters of the underlying Beta distribution, . If the underlying Beta distribution has low variance, it results in a behavior similar to a geometric random variable. Otherwise the variability of probabilities of success alone can generate very heavy tailed behavior. We have generated Monte Carlo samples of size 10^6 of Beta-Geometric random variables. As can be seen in Figure 16, which represents the Maximum to Sum ratios for a Beta-Geometric with $\alpha = 1$ and $\beta = 3$ the mean and higher moments do not converge. Figure 19 presents the Mean Excess function for a Beta-Geometric with $\alpha = 1$ and $\beta = 3$ (blue points), $\alpha = 3$ and $\beta = 1$ (red points), and $\alpha = .5$ and $\beta = 1$ (green points). We can observe, the strictly increasing nature of these functions.

Moreover, a log-transformed Pareto random variable is notoriously exponentially distributed. Hence a comparison of the theoretical quantiles of an Exponential random variable with those of the log-transform of empirical data is the basis of a visual Pareto test, known as the QQ (Quantile-Quantile) plot, Cf. Section 4.1. of [56]. A linear pattern indicates that the empirical data belongs to a generalized Pareto distribution and confirms heavy tails. Figure 20 presents a QQ plot for a 10^6 Monte Carlo sample of a Beta-Geometric with $\alpha = 1$ and $\beta = 3$, in which linearity can be observed. In addition to the previous remarks concerning the non-convergence of moments and the increasing Mean Excess function, this observation makes us conclude that the Beta-Geometric can exhibit the behavior of a generalized Pareto model [53].

Though the main focus of this paper does not lie in a theoretical study of random variables, it can be empirically verified that the tail of this random variable is at least heavy enough to belong to the subexponential class for some values of α and β that result in even thinner tails than the values that seem adequate to model LOS in the dataset at hand. We have generated two random samples of 10^6 observations each from two identical Beta-Geometric distributions with

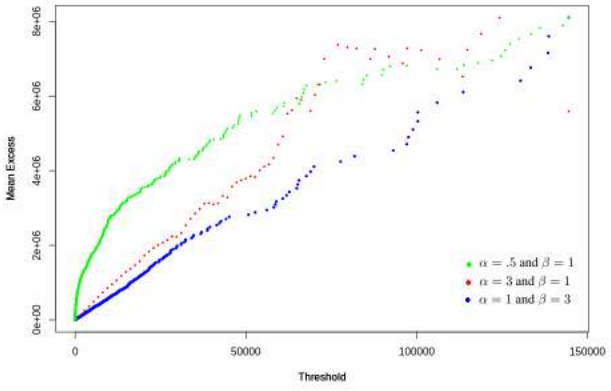

Figure 19: Mean Excess Functions for Beta-Geometric distributions

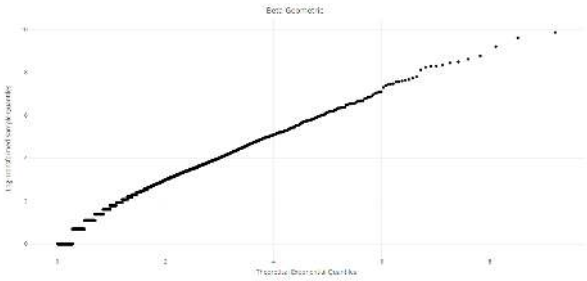

Figure 20: Q-Q plot for a Beta-Geometric

$\alpha \geq 1.1$ and $\beta \geq \alpha$ and verified the property defining subexponentiality. Figure 21 presents the empirical limit of the ratio $\frac{1-F^{(2)}(x)}{1-F(x)}$ for $\alpha = 1.1$ and $\beta \in \{3, 5, 7, 9\}$. As we will show in the subsequent sections, LOS can exhibit even heavier tails than these Monte Carlo samples.

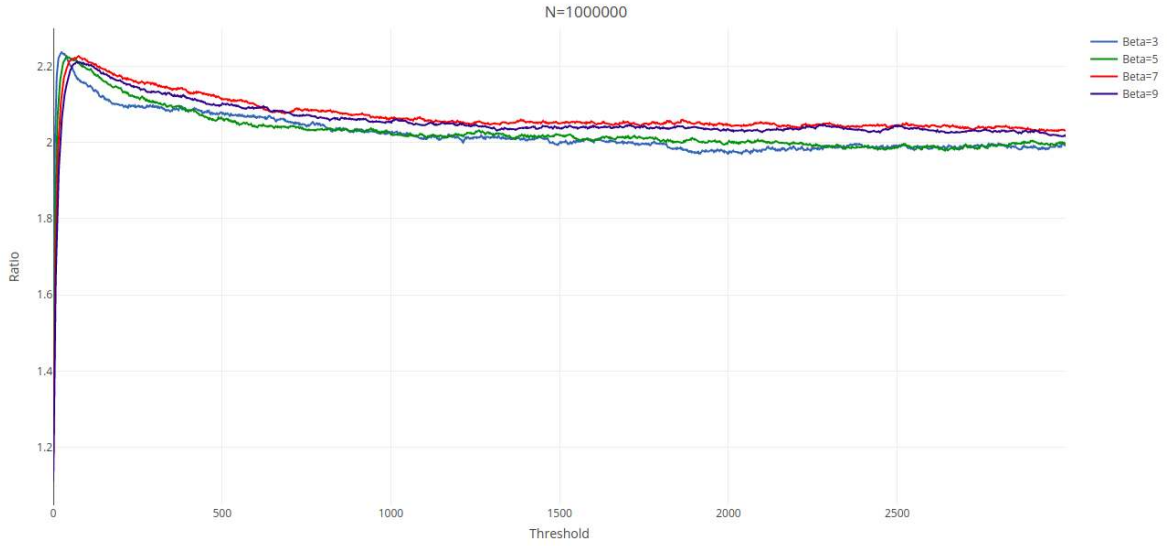

Figure 21: $\frac{1-F^{(2)}(x)}{1-F(x)}$ for a Beta-Geometric with $\alpha = 1.1$ and $\beta \in \{3, 5, 7, 9\}$

5 Results: The role of Extreme LOS Value

This section presents some properties of the empirical LOS in the four specialty departments which are direct consequences of the high frequency and statistical significance of extreme values. We subsequently propose a Beta-Geometric model of LOS and evaluates its goodness of fit.

5.1 On the concentration of LOS

Heavy-tailed distributions tend to be dominated by a small percentage of observations. For LOS, this means that days of hospital would be concentrated among a small percentage of admissions or patients. Such a Power Law behavior which has been previously observed in the English healthcare system in [61]. However, this effect cannot be adequately modeled by thin-tailed random variables, discarding outliers, or considering small samples.

It should be noted that this behavior is a natural consequence of the variability in patients' daily probabilities of discharge. An equal distribution of LOS is neither natural nor desirable. However, from a resource planning point of view, measuring this inequality is important in understanding and modeling healthcare systems in order to adequately price LOS. This is particularly true for high-demand hospitals operating close capacity such as Siriraj hospital, where patients are directly competing for admission.

Figures 26, 27, 28, and 29 represent the reversed Lorenz curves for LOS, by admission and by patient, in Surgery, OB/GYN, Pediatrics, and Rehabilitation Medicine respectively. These plots are Lorenz curves in which the horizontal axis is inverted. Thus, the horizontal axis represents the cumulative percentage of admissions/patients ordered by decreasing LOS, and the vertical axis represents the corresponding cumulative percentage of LOS consumed by that percentage of admissions/patients. These plots read for $x\%$ of patients/admissions have consumed $y\%$ of the LOS. In a thin-tailed distributions, though mildly concave, these plots would typically not show points of inflexion [55], from which the slope of the trendline markedly changes. These points, known as elbows, can serve as a basis for patient classification.

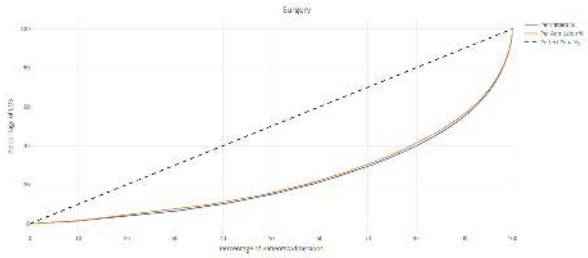

Figure 22: Lorenz curve of LOS in Surgery

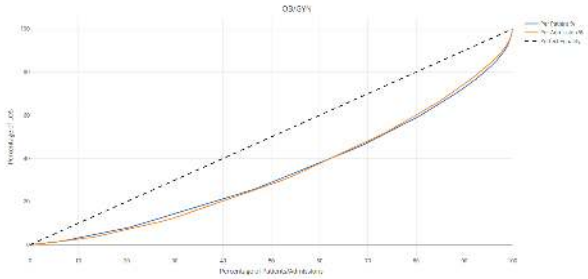

Figure 23: Lorenz curve of LOS in Obstetrics and Gynaecology

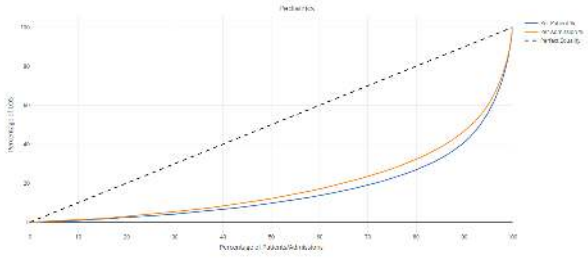

Figure 24: Lorenz curve of LOS in Pediatrics

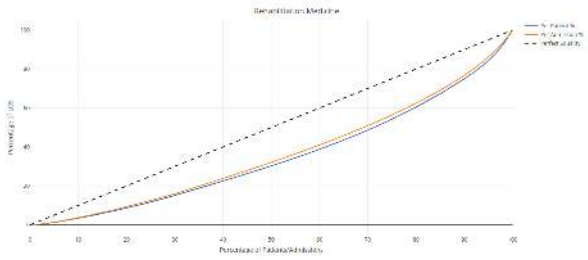

Figure 25: Lorenz curve of LOS in Rehabilitation Medicine

Table 3 presents the Gini index for the LOS per admission and per patient, for each specialty department.

As expected, we can observe that the heavier the tail of the distribution, the higher the inequality in the distribution of LOS, and the aggregation of LOS per patient tends to make the tails significantly heavier.

The distribution of LOS in Rehabiliation Medicine, once again, exhibits a behavior that is the closest to that of a thin-tailed random variable. However, we note an inflexion point at the left of the curve, where 3.05% of admissions concentrate 12.45% of LOS, and 3.6% of patients concentrate 12.36% of LOS. OB/GYN presents a similar but more extreme inflexion point, where 1.94% of admissions concentrate 10% of LOS and 2.58% of admissions concentrate 12.79%. Surgery and Pediatrics exhibit the most extreme discrepancies in the distribution of LOS between the top percentiles and the rest of the admissions/patients, thus resulting in lower bed turnover rates. In Surgery, 4.2% of

Statistic	Surgery	OB/GYN	Pediatrics	Reh. Med.
Nº of Admissions	17949	19922	7499	994
Gini index	.53	.32	.62	.27
Delta	7.07	2.51	10.29	13.47
Nº of Patients	14884	17483	5086	304
Gini index	.55	.32	.67	.30
Delta	8.78	2.84	16.44	48.59

Table 3: Gini index for LOS per admission and per patient in the four departments

admissions/patients concentrate 30% of LOS. and in Pediatrics 6.7% of patients concentrate 50% of LOS, and 5.5% of admissions concentrate 40% of LOS. The latter specialty department also exhibits the widest gaps between the curves per admission and per patient, which illustrates the effect of multiple admissions on the concentration of LOS. The inflexion points determined by the percentile distribution of LOS allow for the definition of these thresholds in a way that takes actual resource consumption in a care unit into account.

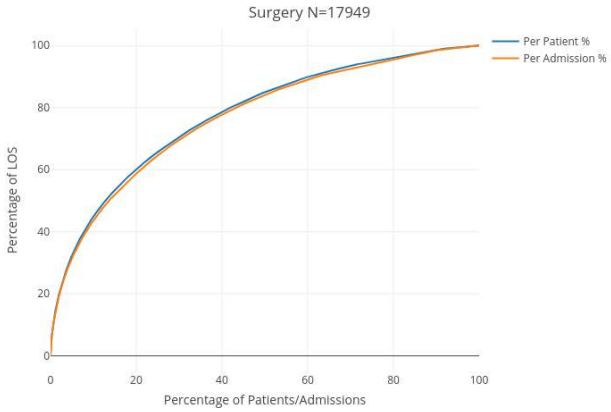

Figure 26: Reversed Lorenz curve of LOS in Surgery

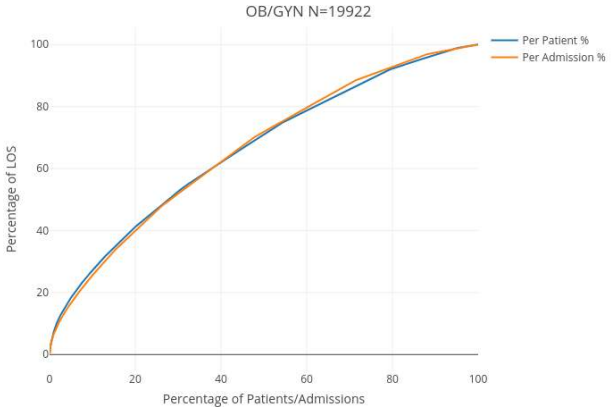

Figure 27: Reversed Lorenz curve of LOS in Obstetrics and Gynaecology

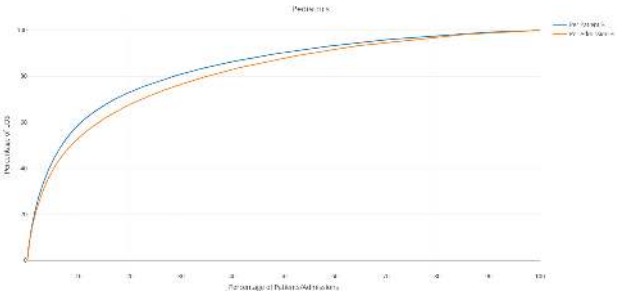

Figure 28: Reversed Lorenz curve of LOS in Pediatrics

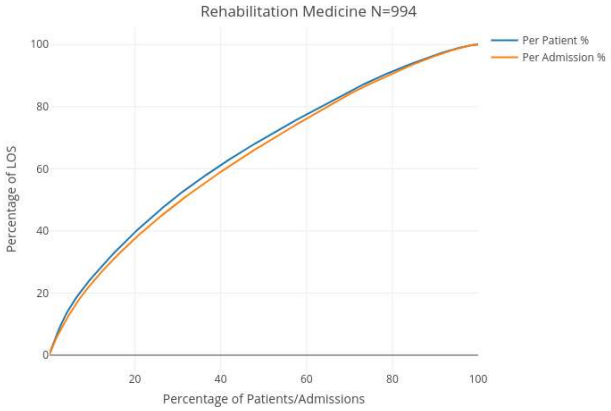

Figure 29: Reversed Lorenz curve of LOS in Rehabilitation Medicine

5.2 On revenue

As seen in Section 5.1, the top percentiles of patients/admissions, in terms of LOS, can represent a very significant proportion of total LOS. This concentration of LOS also has financial effects. In Figures 30, 31, 32, and 33 we have

computed the average daily charge per admissions, as a function of LOS (i.e. revenue divided by LOS, for each discrete value of LOS), in Surgery, OB/GYN, Pediatics, and Rehabilitation Medicine, respectively.

For the three departments of Surgery, OB/GYN and Pediatrics, we find that *ceteris paribus*, the mean revenue per day of an admission significantly decreases with the increase of LOS. That is up to a certain threshold, from which long LOS patients are approximately charged the same amount per day. Rehabilitation Medicine is once again a singular specialty department in this regards. Indeed, we find that the charge per day tends to increase with the increase of LOS for short stay patients, up to a point (from which long LOS could be defined) where it stabilizes for additional days of stays. The curves in Figures 30 to 33 can be accurately described as piecewise linear. Using the Elbow method [55], the respective thresholds for the pseudo-linearity of the charge per day function in each of the four specialty department are given in Table 4.

In Figure 30, 31, and 32 we observe a steeply declining, pseudo-linear curve up to a certain threshold of LOS, which could be considered the threshold for short LOS in terms of charge. For this group of patients, additional days of LOS result in a decrease in the charge per day (in addition to the opportunity cost resulting from the inability to admit patients with a shorter stay and thus higher charge). Once the function the charge per day function reaches its minimum, it stabilizes for all higher LOS values.

Department	Surgery	OB/GYN	Pediatrics	Reh. Med.
First Range	1 to 14 days	1 to 11 days	1 to 13 days	1 to 9 days
Correlation	$-.94$	$-.88$	$-.91$	$+.83$
Second Range	15 to 30 days	12 to 17 days	14 to 50 days	10 to 39 days
Correlation	$-.90$	$-.90$	$-.91$	$+.73$
Third Range	31 to 304 days	18 to 119 days	51 to 272 days	40 to 87 days
Correlation	$-.95$	$-.94$	$-.97$	$+.90$

Table 4: Thresholds and correlation for the average charge per day of LOS

We find a very high negative piecewise linear correlation ($\leq -.88$) for each pseudo-linear range of the average charge per day function in Surgery, Ob/GYN, and Pediatrics, as detailed in Table 4. This hints at the fact that long LOS is not adequately priced, possibly because long LOS observations are discarded as outliers in analysis models. To give a prosaic example, let us focus on one hospital bed and compare either admitting n patients with 1 days of LOS or one patient with n days of LOS. Additionally, we assume that the average charge per day for a stay of 1 day is m monetary units, when the average charge per day for a stay of n days is 1 monetary unit. Not only would the single patient staying n days pay less per day ($m > 1$), but the difference in charge with shorter stay patients would be multiplied by n such that the total opportunity cost would be $n \cdot (m - 1)$. This in addition to the fact that this hospital would have been used to serve $n - 1$ more persons. This simplistic example obviously neglects the medical needs of these patients and should not be seen as a call to make admission decisions based on revenue. It is however an invitation to consider renegotiating the pricing strategy at Siriraj Hospital in order to adequately price the opportunity cost resulting from long LOS. However, the Rehabilitation Medicine department seems to adequately price this opportunity cost (possibly because long LOS is expected in this department) as illustrated in Figure 33. The average charge per day increases with LOS, and shows a similarly high positive piecewise linear correlation when broken down into three pseudo-linear ranges. Similarly to the CDF (Figure 4), the average charge function in this specialty department is sigmoidal. We note that a change in the nature of the function from convex to concave occurs precisely at the 20 days mark. This a reflection of the mixed nature of the distribution of LOS in Rehabilitation Medicine, with a marked difference in the distribution of values of LOS below and above 20 days

Another question concerns the association between LOS, charge, and the discharge status of patients (broken down into positive and negative statuses, as described in Section 3). Figures 34, 36, and 38 show individual LOS and charges per day of admissions concluded with a positive discharge status in Surgery, OB/GYN, and Pediatrics, respectively, while Figures 35, 37, and 39 represents these variables for admissions concluded with a negative discharge status for the same respective departments. The Rehabilitation Medicine department having seen only a single admission with a negative discharge status ("Not improved"), the same data is represented in Figure 40 and is excluded from the following remark. Although discharges with a negative discharge status are much less frequent, we observe a similar distribution of LOS in each of the positive and negative discharge status group, as with the distribution of LOS over both groups.

Moreover, the distribution of LOS over each group shows a similar correlation with the average charge per day as the distribution over the whole. Overall, we can conclude that LOS, at this level of analysis (whole medical specialty departments) does not show any striking association with the discharge status. It may however possibly be the case of a more granular analysis per DRG, which is outside the scope of the present study.

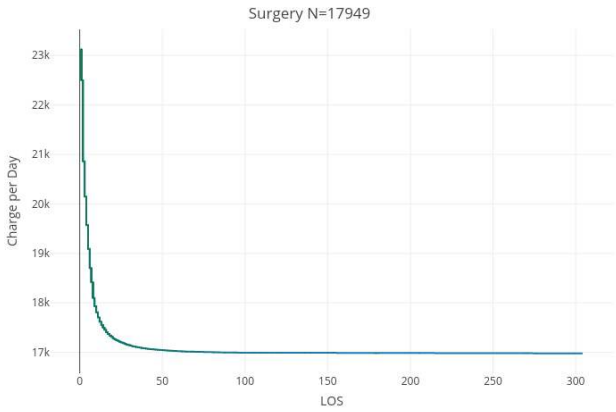

Figure 30: Average charge per day in Surgery

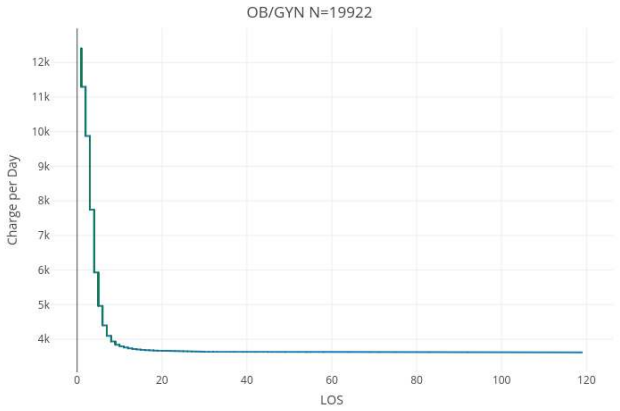

Figure 31: Average charge per day in Obstetrics and Gynaecology

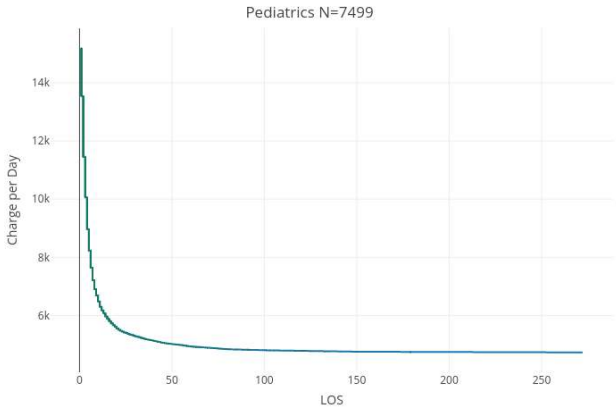

Figure 32: Average charge per day in Pediatrics

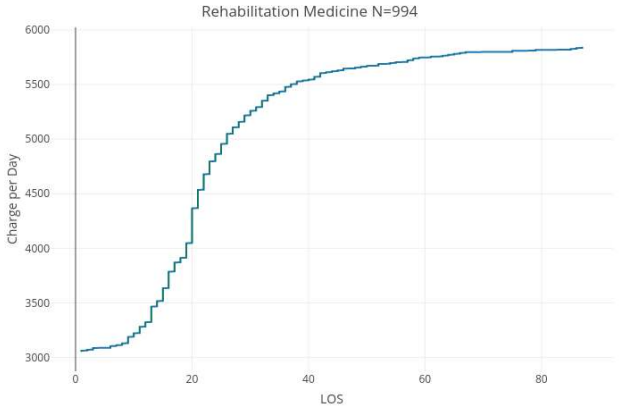

Figure 33: Average charge per day in Rehabilitation Medicine

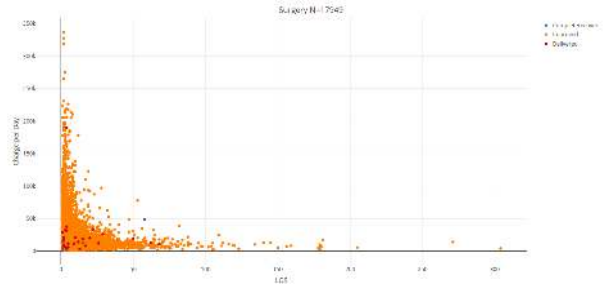

Figure 34: Charge per day for positive discharge status in Surgery

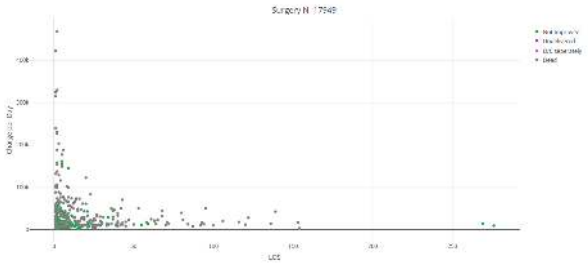

Figure 35: Charge per day for negative discharge status in Surgery

5.3 On the Convergence of Moments

Because of the preponderance of extreme values, one of the main concerns when suspecting subexponentionality in the study of random variables concerns the existence of moments, as described in Section 4.2. Indeed, the non-convergence of moments, or their very slow convergence (which requires extremely large sample sizes), renders the estimation of indicators such as the mean, variance, skewness or kurtosis from empirical observations impractical. Figures 41, 42

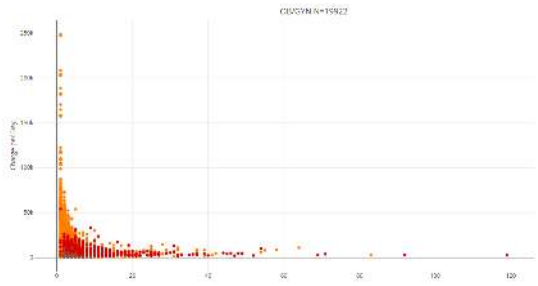

Figure 36: Charge per day for positive discharge status in Obstetrics and Gynaecology

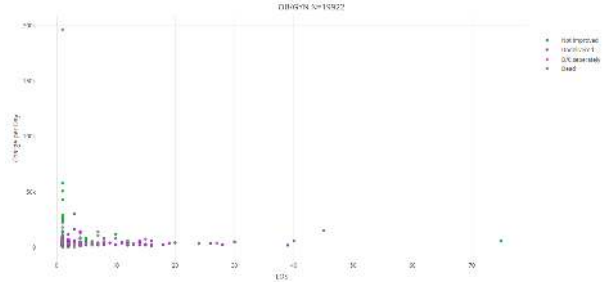

Figure 37: Charge per day for negative discharge status in Obstetrics and Gynaecology

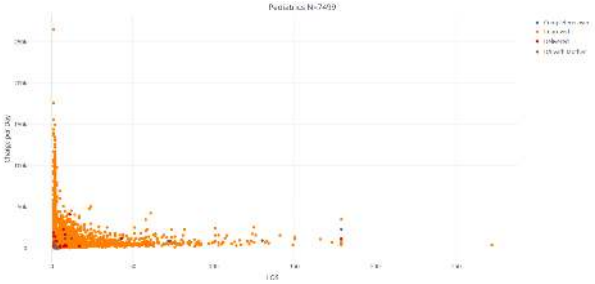

Figure 38: Charge per day for positive discharge status in Pediatrics

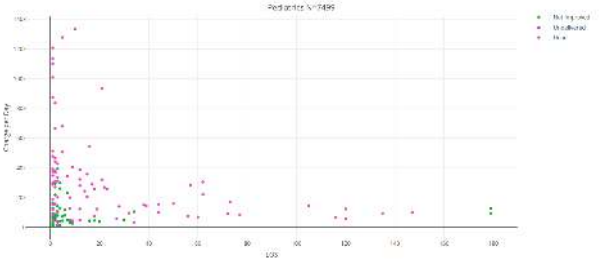

Figure 39: Charge per day for negative discharge status in Pediatrics

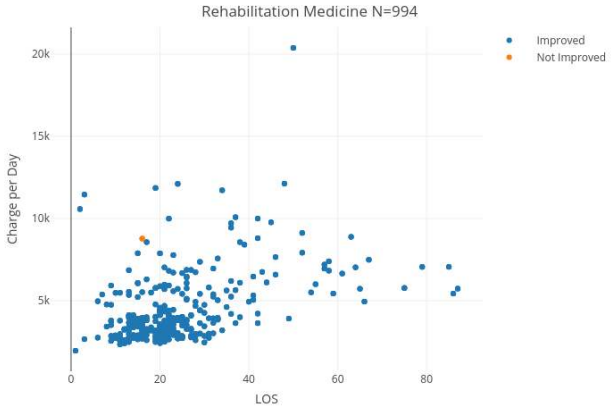

Figure 40: Charge per day in Rehabilitation Medicine

represent the Maximum to Sum plot for LOS per admission and per patient, respectively, in Surgery. Figures 43 and 44, 45 and 46, as well as 47 and 48 represent similar pairs of Maximum to Sum plots for OB/GYN, Pediatrics, and Rehabilitation Medicine. We note the convergence of first moments (mean) in all specialty departments. However, second moments (variance), though apparently convergent, show a slow convergence even for our large datasets of tens of thousands of patients, which makes empirical estimates unreliable. Similarly, third and fourth moments (skewness and kurtosis respectively) cannot be reliably estimated from empirical data, except in the Rehabilitation Medicine department with per admission data.

This problem is particularly acute for LOS data per patient in Surgery and Pediatrics. As previously noted, the aggregation of LOS per patient makes the tail of the distributions of LOS heavier and the convergence of moments accordingly slower.

The novel information from these plots is the visible heavy tailedness of the distribution of LOS per patient in Rehabilitation Medicine.

We have excluded a Generalized Pareto distribution for each specialty department using the Pareto test included in R Package *ptsuite* [62, 63]. The resulting p-values were respectively $4.52 \cdot 10^{-193}$, 0, $1.24 \cdot 10^{-291}$, and 0 in Surgery, OB/GYN, Pediatrics, and Rehabilitation Medicine and are not consistent with generalized Pareto distributions. However, the Maximum to Sum plots are consistent with subexponential distributions, within which the Beta-Geometric is the best candidate because of its discreteness and consistency of assumptions in modeling LOS, as stated in Section 4.5. This is further confirmed with QQ-plots and Mean Excess plots.

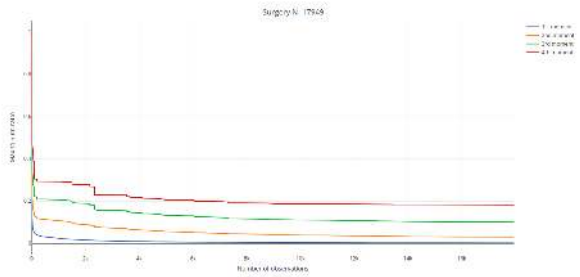

Figure 41: Maximum to Sum ratios of LOS per admission in Surgery

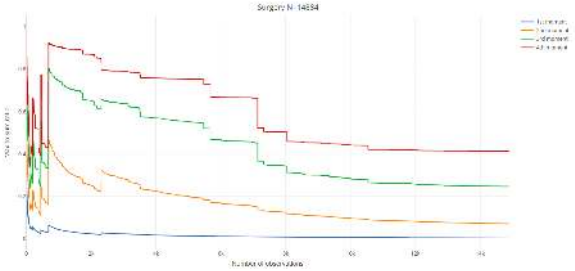

Figure 42: Maximum to Sum ratios of LOS per patient in Surgery

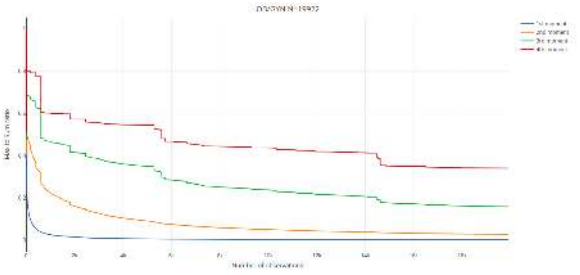

Figure 43: Maximum to Sum ratios of LOS per admission in Obstetrics and Gynaecology

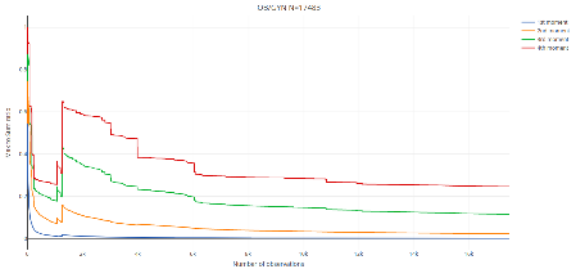

Figure 44: Maximum to Sum ratios of LOS per patient in Obstetrics and Gynaecology

Indeed, in Figures 49 to 56, representing the QQ-plots of LOS per admission and per patient, in the four department, we can observe a pseudo-linear behavior (including in Rehabilitation Medicine, past a certain threshold) that is consistent with a subexponential distribution. However, the marked difference between the head and tail of the QQ-plots in Figure 55 and 56 indicates that the distribution of LOS in Rehabilitation Medicine being possibly of a mixed nature, which is further confirmed by the Mean Excess plots in the next Section.

5.4 On Expected Residual Length of Stay

An increasing expected residual LOS is possibly the most important property that cannot be captured by thin-tailed models. As explained in Section 4.3, this property can be understood as the expectation of LOS increasing the longer a patient has been in the system, or prosaically, the longer a patient has been admitted, the further they, counter-intuitively, get from being discharged in subsequent days.

Figures 57 to 60 represent the mean excess plots for LOS per admission and per patient, in Surgery, OB/GYN, and Pediatrics. They confirm the subexponential nature of these variables. Note that the decrease in the Mean Excess function at the rightmost points of the curve is a common effect known as the "finite sample bias" [40], wherein points close to the sample maximum would not be expected to further increase simply because no higher values have been

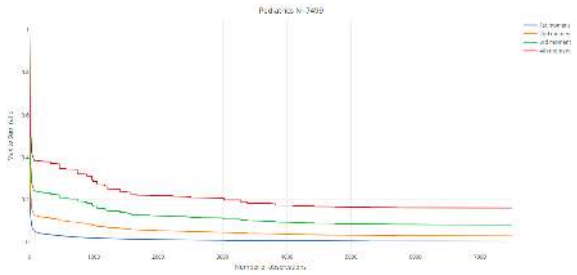

Figure 45: Maximum to Sum ratios of LOS per admission in Pediatrics

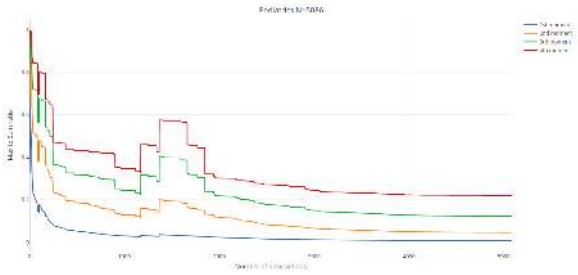

Figure 46: Maximum to Sum ratios of LOS per patient in Pediatrics

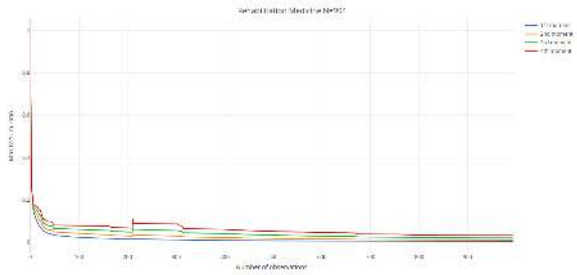

Figure 47: Maximum to Sum ratios of LOS per admission in Rehabilitation Medicine

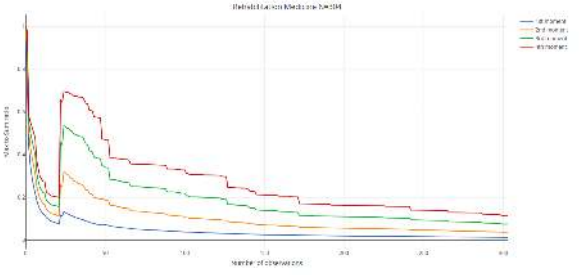

Figure 48: Maximum to Sum ratios of LOS per patient in Rehabilitation Medicine

observed yet.
The Mean Excess function of LOS in Rehabilitation Medicine, which is represented per admission in Figure 63, and per patient in Figure 64, once again stands out. The Mean Excess decreases for thresholds below 20 We see a dramatic change in the monotonicity of the Mean Excess Function, which goes from decreasing to increasing precisely at value 20 days. This indicates a heavy-tailed distribution for stays above 20 days.

Lastly, we observe that all increasing Mean Excess functions exhibit more linear behavior when LOS is aggregated by patient. This confirms, once again, the increase in the heaviness of tails that this aggregation induces.

6 Modeling Length of Stay

Given n independent realizations $x_1, \ldots, x_n$, of a Beta-Geometric random variable X, defined by parameters α and β as described in Section 4.5, the likelihood function is given by the product of their individual PDF:

$$L = \prod_{i=1}^{n} \frac{B(\alpha + 1, \beta + x_i - 1)}{B(\alpha, \beta)}$$

The corresponding log-likelihood is thus given by:

$$Log(L) = Log(B(\alpha + 1, \beta + x_i - 1)) - nLog(B(\alpha, \beta))$$

The maximum likelihood estimates for α and β, respectively $\hat{\alpha}$ and $\hat{\beta}$, can be obtained by numerically maximizing $Log(L)$ with respect to α and β using statistical software such as the *betageometric* function of R package VGAM [32] or the *BGEPDF* of statistical software *dataplot* [35, 38, 39].

Based on the evidence for a Beta-Geometric distribution of LOS in these departments, we have computed the maximum likelihood estimates of the α and β parameters for LOS in Pediatrics, Surgery, and OB/GYN.

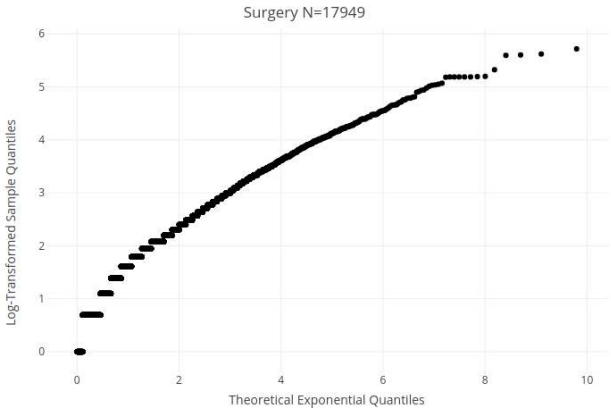

Figure 49: Q-Q plot of LOS per admission in Surgery

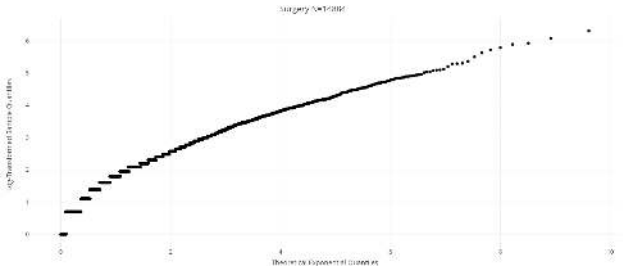

Figure 50: Q-Q plot of LOS per patient in Surgery

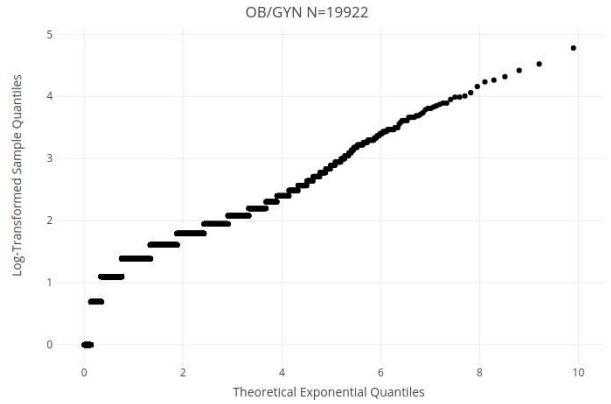

Figure 51: Q-Q plot of LOS per admission in Obstetrics and Gynaecology

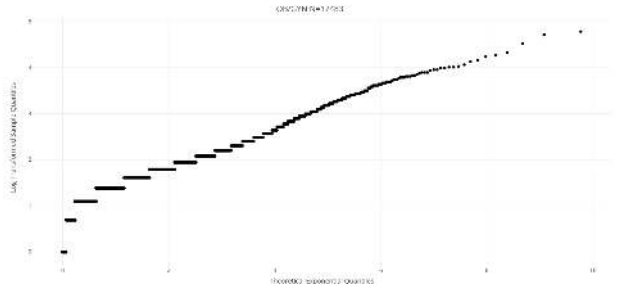

Figure 52: Q-Q plot of LOS per patient in Obstetrics and Gynaecology

As for the Rehabiliation Medicine department, and given the marked difference in the distribution before and after 20 days, and the evidence for a mixture distribution, we fit a Poisson distribution for values of LOS from 1 to 19 days with a Poisson distribution, whereas the distribution of values above 20 days is modeled with a shifted Beta-Geometric. We subtract 20 days from all observations, fit a Beta-Geometric, then add 20 days to the resulting Beta-Geometric distribution.

We generate a 10^5 Monte-Carlo sample for the fitted distribution to the LOS in the four departments, both per admission and per patient. To evaluate goodness of fit, we use a discrete two-sample Kolmogorov–Smirnov goodness of fit test [46, 47, 48].

The Kolmogorov–Smirnov statistic (the D-statistic) more precisely quantifies the distance between the observed distribution function of the sample and the cumulative distribution function of the fitted distribution. We present the α and β parameters found for the fitted Beta-Geometric distributions of LOS in the four departments[1], as well as the corresponding Kolmogorov–Smirnov distance, at 5% level of significance, in Table 5.

Moreover, Figures 65 and 66 respectively compare the histograms of the fitted and observed LOS per admission in Pediatrics and Rehabilitation medicine. Figures 67 and 68 respectively illustrate the comparison of cumulative distributions functions per patient in Surgery and per admission in Rehabilitation Medicine, and Figures 69 and 70 the comparison of Mean Excess Functions for the distribution of LOS per patient in Surgery and per admission in Rehabilitation Medicine.

[1]For the Rehabilitation Medicine department this distribution only concerns values of LOS above 20 days as previously stated.

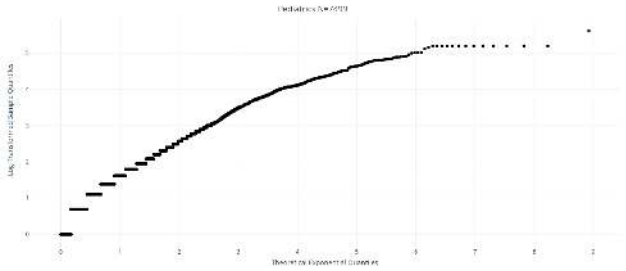

Figure 53: Q-Q plot of LOS per admission in Pediatrics

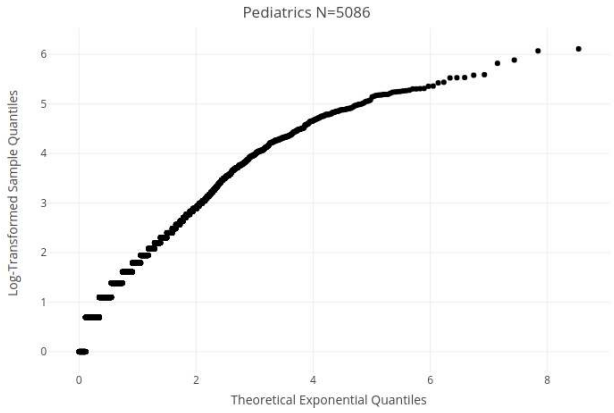

Figure 54: Q-Q plot of LOS per patient in Pediatrics

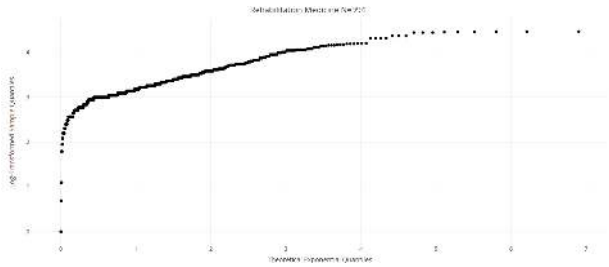

Figure 55: Q-Q plot of LOS per admission in Rehabilitation Medicine

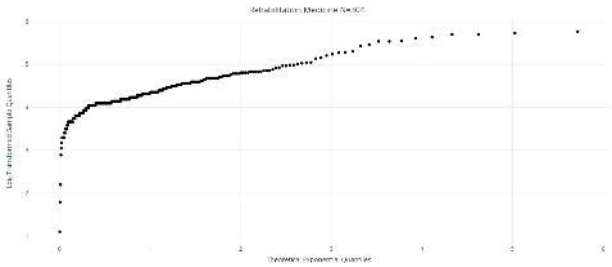

Figure 56: Q-Q plot of LOS per patient in Rehabilitation Medicine

Fitting mixture distributions is a relatively under-studied problem in statistics, and overwhelmingly concerned with Gaussian mixtures [49, 50]. To the best of our knowledge there haven't been any works regarding fitting a Poisson/Beta-Geometric mixture. As a result, fitting a theoretical distribution to the LOS in Rehabilitation medicine proved more challenging because of the mixed nature of the distribution of LOS in this department resulting from the artificial inflation of the frequency of value 20 days. However, for the three departments whose LOS only sees unconstrained variations (Surgery, Pediatrics, and OB/GYN), we obtain a satisfying fit with Beta-Geometric models and a 95% confidence level, as described in Table 5.

Model	Surgery	OB/GYN	Pediatrics	Reh. Med.
Admissions	17949	19922	7499	994
α	7.027677	7.1307	3.2695	37.3032
β	38.58531	9.3826	17.6431	387.4253
D-statistic	0.0480	0.0398	0.0474	0.0605
Patients	14884	17483	5086	304
α	6.1900	13.9223	2.3282	10.4402
β	39.6485	4.0856	14.9977	24.5310
D-statistic	0.0590	0.0483	0.0518	0.0632

Table 5: Results of the Kolmogorov-Smirnov tests for fitted 10^5 Beta-Geometric samples with 95% confidence level. A Poisson model was additionally used for stays below 20 days in Rehabilitation Medicine.

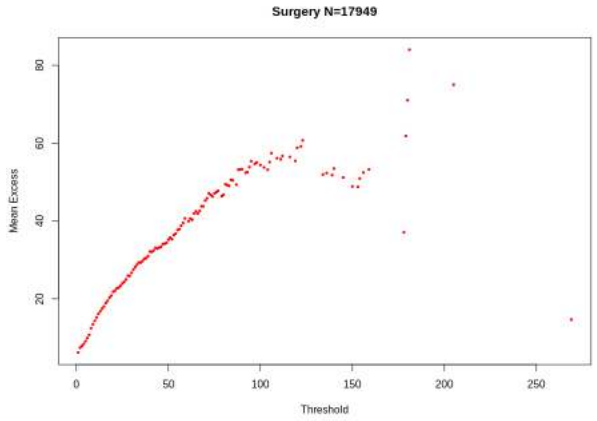

Figure 57: Mean Excess function of LOS per admission in Surgery

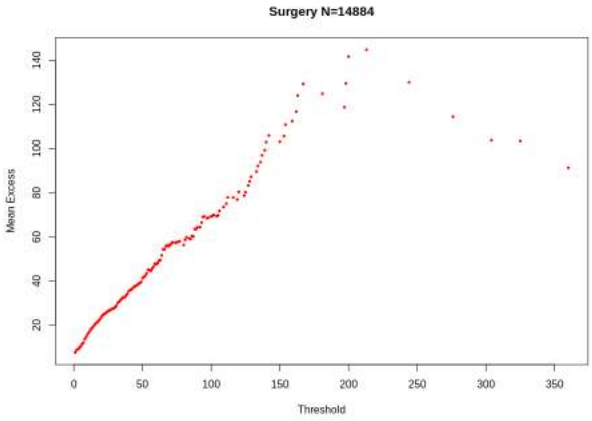

Figure 58: Mean Excess function of LOS per patient in Surgery

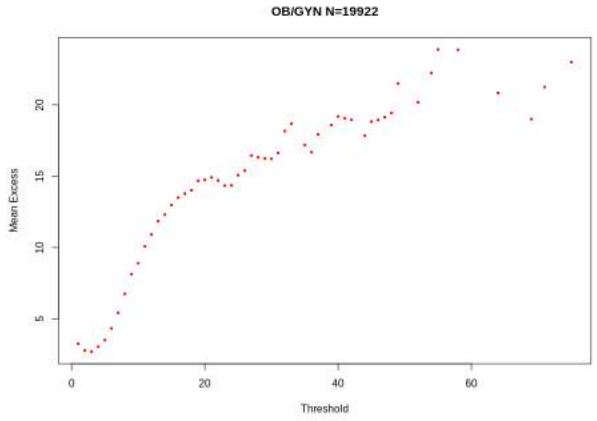

Figure 59: Mean Excess function of LOS per admission in Obstetrics and Gynaecology

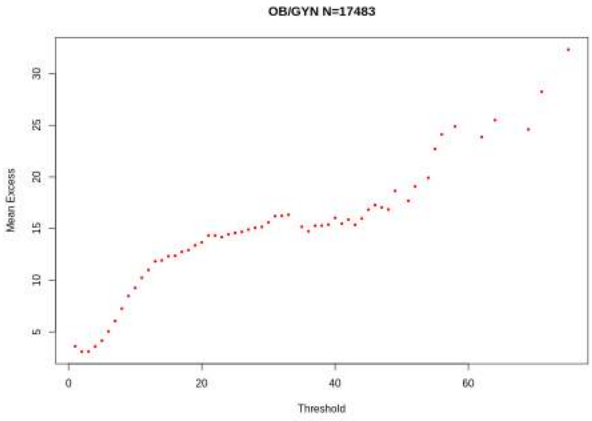

Figure 60: Mean Excess function of LOS per patient in Obstetrics and Gynaecology

7 Managerial Implications

Section 5 identified and analyzed some empirical properties of LOS that have important implications for operational and financial planning at the level of healthcare units.

We have shown, in Section 4.1, that extreme value can result in large discrepancies in the distributions of LOS among admissions/patients.

Moreover, a counter-intuitive property of LOS studied in Section 5.4 is an increasing conditional expectation, as represented by the shape of the mean excess functions in Figures 57 to 60. This is the case for all departments, including the Rehabilitation Medicine department past a certain threshold, as seen in Figures 63 and 64. This property makes Length of Stay Lindy [24], in an undesirable sense; the longer a patient has stayed, the longer he/she would be expected to remain, with a corresponding increase in the expected opportunity cost of high LOS.

Thus, from the standpoints of both equity of access to healthcare and adequately reflecting the value offered to patients, the opportunity cost of long stays should be reflected in their pricing. Section 5.2 showed that this opportunity cost is currently not adequately taken into account at Siriraj hospital. Models of Revenue Management based on differentiated pricing models, which are now standard in the Airline Industry [64], could be fruitfully applied by healthcare providers to cover this opportunity cost and provide beds to patients who value them most when demand exceeds supply.

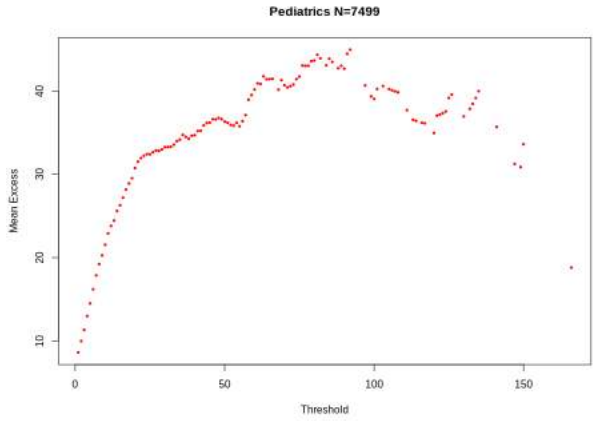

Figure 61: Mean Excess function of LOS per admission in Pediatrics

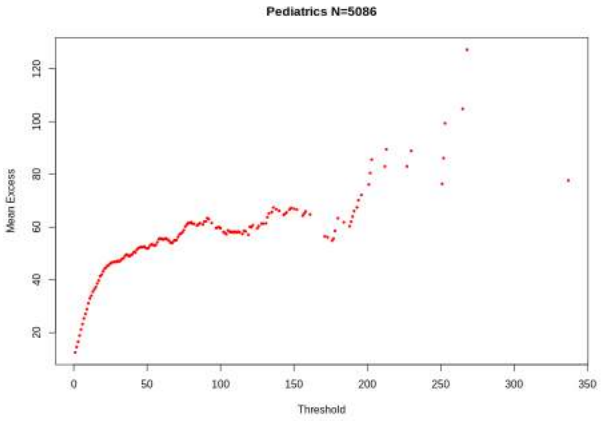

Figure 62: Mean Excess function of LOS per patient in Pediatrics

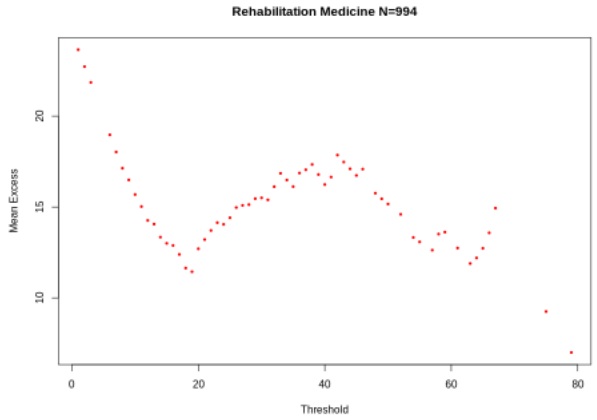

Figure 63: Mean Excess function of LOS per admission in Rehabilitation Medicine

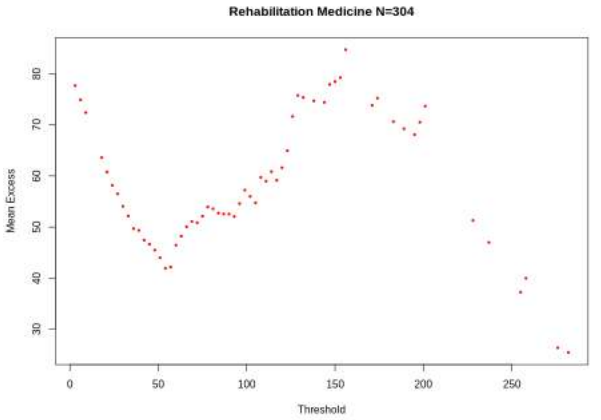

Figure 64: Mean Excess function of LOS per patient in Rehabilitation Medicine

Further, in Section 5.3, we have highlighted some practical difficulties in estimating the moments of LOS from empirical observations due to the slow convergence of the Law of Large Numbers. This *catastrophe principle*, inherent to subexponential random variables, makes models for forecasting precise values of LOS (e.g. least square regressions based on second moments) inapplicable. Lastly, we have illustrated the importance of reducing unnecessary readmissions and avoiding multiple admissions. As they compound to make the tails heavier, they exaggerate the previous challenges.

Based on the simple assumptions stated in 4.5, we have shown, in Section 6, that seeing hospital admissions as a simple Geometric process with heterogeneous (Beta) probabilities of discharge adequately reproduces these properties and provides a good fit with observed LOS.

Rather than a naive forecasting of length of stay at admission, only evaluated by absolute percentages of deviations or correlation, a more actionable quantitative focus for resource planning lies in accurately estimating the parameters of the Beta-Geometric model that governs a given LOS. This type of analysis has the additional advantage of being scalable to different levels of decision-making (i.e. variations in the distribution of the discharge probability would result in different Beta-Geometric models of LOS in a hospital, departments, units, and within those for different DRG, type of patients, types of interventions, etc.). Further, the various graphical tools we have proposed could be employed to dynamically determine thresholds and use them for long-stay patient reviews. Moreover, pricing negotiations with

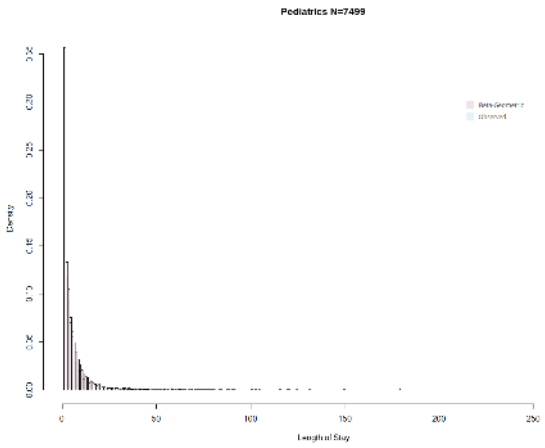

Figure 65: Histogram of observed and fitted LOS per admission in Pediatrics

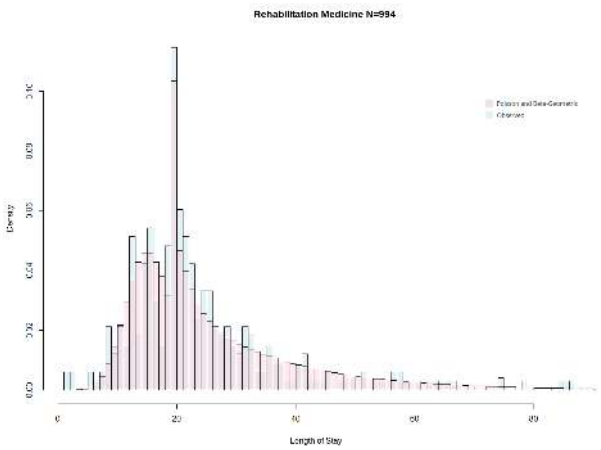

Figure 66: Histogram of observed and fitted LOS per admission in Rehabilitation Medicine

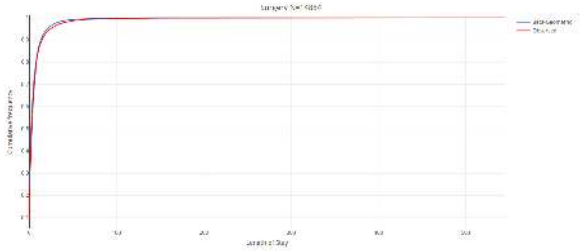

Figure 67: CDF of observed and fitted LOS per patient in Surgery

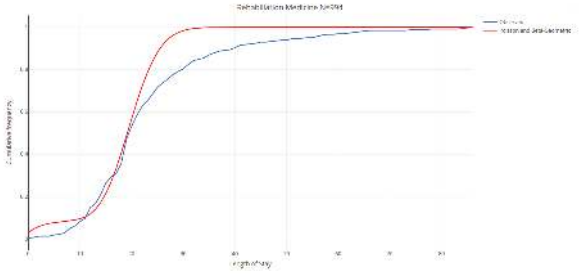

Figure 68: CDF of observed and fitted LOS per admission in Rehabilitation Medicine

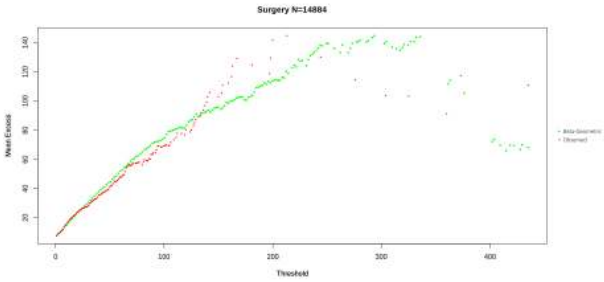

Figure 69: Mean Excess function of observed and fitted LOS per patient in Surgery

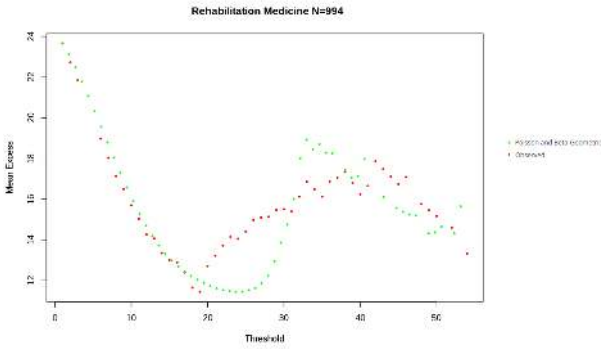

Figure 70: Mean Excess function of observed and fitted LOS per admission in Rehabilitation Medicine

payers should not only include the direct cost of treatment but also the opportunity cost of the expected LOS increasing in the tail, as well as the opportunity cost of being unable to admit patients with potentially shorter stays and higher charges per day, as a result of tail events.

8 Conclusions and Limitations

This study showed that long stays (the tail of LOS distributions) have important consequences for resource consumption and revenue management in healthcare facilities and are not to be discarded as outliers. It is thus important that healthcare providers move away from simplistic, thin-tailed models and the disproportionate focus on regression, to align their quantitative models with those of payers (Extreme Value Theory is, after all, the mathematics of insurance). We have proposed a Beta-Geometric model for LOS that adequately reproduces these properties and shows a satisfying fit with empirically observed LOS in $46,364$ electronic health records, covering four specialty departments. An added advantage of the proposed model is that it offers a consistent model for decision-support in a medical specialty department that can also be scaled down to individual DRG, or up the hospital as whole, by a simple adjustment of the parameters of the underlying Beta distribution of discharge probabilities. Moreover, the discreteness and simplicity of the assumptions that this model rests on offer an advantage over mixture distribution models. By Occam's razor these models are not justified when a single (compound) probability distribution can adequately fit. However, we have found that artificial restrictions on LOS (such as the preponderance of LOS value 20 in Rehabiliation Medicine) can produce emprical LOS distributions that are best described by mixture models. In these cases, modeling with a Beta-Geometric/Poisson mixture distribution proved more challenging and produced a lower quality fit. A more refined analysis of LOS in Rehabilitation Medicine, over larger datasets than the one considered in our study (994 admissions), and with additional methodological developments on fitting mixed distributions would be potentially valuable.

References

[1] Roberts, R. R., Frutos, P. W., Ciavarella, G. C., et al. Distribution of fixed versus variable costs of hospital care. Journal of the American Medical Association, 281:644–9, 1999.

[2] Faddy, M., Graves, N., Pettitt, A. Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. Value in Health. 12(2):309-314, 2009.

[3] Taleb, N. N., Bar-Yam, Y., Cirillo, P. On Single Point Forecasts for Fat-Tailed Variables. International Journal of Forecasting, in press, available online, 2020.

[4] Davidson, J. Statistical Analysis of the Regression Model. Econometric Theory. Oxford: Blackwell. pp. 17–36, 2000.

[5] Marazzi, A., Paccaud, F., Ruffieux, C., Beguin, C. Fitting the Distributions of Length of Stay by Parametric Models. Medical Care, 36(6):915–927, 1998.

[6] Behboodian, J. On the modes of a mixture of two normal distributions. Technometrics, 12:131–139. doi:10.2307/1267357. JSTOR 1267357, 1970.

[7] Lee, A. H., Fung, W. K., Fu, B, Analyzing Hospital Length of Stay Mean or Median Regression? Medical care, 41(5): 681–686, 2003.

[8] Sills, M. R., Huang, Z. J., Shao, C., et al. Pediatric Milliman and Robertson length-of-stay criteria: Are they realistic? Pediatrics, 105:733–737, 2000.

[9] Leung, K. M., Elashoff, R. M., Rees, K. S., et al. Hospital- and patient-related characteristics determining maternity length of stay: A hierarchical linear model approach. American Journal of Public Health, 88:377–381, 1998.

[10] Ad, N., Holmes, S.D., Shuman, D.J., et al. Potential impact of modifiable clinical variables on length of stay after first-time cardiac surgery. Annals of Thoracic Surgery, 2015(100):2102–2108, 2015.

[11] Andrei, A.C. Modeling Hospital Length of Stay Data: Pitfalls and Opportunities. Annals of Thoracic Surgery, 2016(01):2425–2432, 2016.

[12] Shahian, D. M., O'Brien, S. M., Filardo, G., et al. Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1—coronary artery bypass grafting surgery. Annals of Thoracic Surgery, 2009.

[13] Pukelsheim, F. The Three Sigma Rule. American Statistician. 48: 88–91. JSTOR 2684253, 1994.

[14] Ickowicz, A., Sparks, R. Modelling hospital length of stay using convolutive mixtures distributions. Statistics in Medicine, 4(1), 2018.

[15] Rady, A. A., El Sayed, S. Investigate the Optimum Model for Length of Stay and Mortality Prediction in the Intensive Care Unit. Journal of Perioperative and Critical Intensive Care Nursing, 4:2, 2018.

[16] Atienza, N., Garcia-Hera, J., Munoz-Pichard, J. M., Villa, R. An application of mixture distributions in modelization of length of hospital stay. Statistics in Medicine, 27(9), 2007.

[17] Gardiner, J. C, Luo, Z., Tang, X., Ramamoorthi, R. V. Fitting Heavy-Tailed Distributions to Health Care Data by Parametric and Bayesian Methods. Journal of Statistical Theory and Practice, 8:4, 2014.

[18] Baek, H., Cho, M., Kim, s., Hwang, H., Song, M., Yoo, S. Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. PLoS One, 13(4), 2018.

[19] Harini S., Subbiah, M., Srinivasan, M. R. Fitting length of stay in hospitals using transformed distributions. Journal of Communications in Statistics: Case Studies, Data Analysis and Applications, 4(1), 2018.

[20] Kernick, D. Wanted—new methodologies for health service research. Is complexity theory the answer?. Family Practice, 23(3), 2000.

[21] Langford, E., Schwertman, N., Owens, M.. Is the property of being positively correlated transitive? The American Statistician, 55, 322-325, 2001.

[22] Pender, J. The truncated normal distribution: Applications to queues with impatient customers, Operations Research Letters, 43, 2015.

[23] Grubbs, F. E. Procedures for detecting outlying observations in samples". Technometrics. 11(1):1–21, 1969.

[24] Taleb, N. N. Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications - Papers and Commentary. STEM Academic Press, 2020.

[25] Teugels, J. L. The class of Subexponential Distributions. The Annals of Probability, 3(6):1000–1011, 1975.

[26] Taleb, N. N. The Black Swan: The Impact of the Highly Improbable, volume 2. Random house, 2007.

[27] Taleb, N. N. Antifragile: Things that gain from disorder. Random House and Penguin, 2013.

[28] Fader, P. S. Hardie, B. G. S. How to Project Customer Retention. Journal of Interactive Marketing, 21(1), 2007.

[29] Johnson, N. L., Kotz, S., Balakrishnan, N. Chapter 21: Beta Distributions, Continuous Univariate Distributions, volume 2. Wiley, 1995.

[30] Dubey, S. D. Compound gamma, beta and F distributions. Metrika. 16:27–31, 1970.

[31] Cyganska, M. The impact factors on the hospital high length of stay outliers. 3rd GLOBAL CONFERENCE on BUSINESS, ECONOMICS, MANAGEMENT and TOURISM, 26-28 November, Rome, Italy, 2015.

[32] Yee, T. W. The VGAM package for categorical data analysis. Journal of Statistical Software, 32:1–34, 2010.

[33] Weinberg, P., Gladen, B.C. The Beta-geometric distribution applied to comparative fecundability studies. Biometrics, 42, 1986.

[34] Singh, B. P., Pudir, P. S., Maheshwari, S. Parameter Estimation of Beta-Geometric Model with Application to Human Fecundability Data, arXiv Statistics/Applications, arXiv:1405.6392, 2014.

[35] Filliben, J. J., Heckert, A. Dataplot Reference Manual: The BGEPDF Library Function. Statistical Engineering Division, National Institute of of Standards and Technology (NIST), US Department of Commerce, 2006. Retrieved from https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/bgepdf.htm

[36] Gastwirth, J. L. The Estimation of the Lorenz Curve and Gini Index. The Review of Economics and Statistics, 54(3):306-316, 1972.

[37] Dunford, R., Su Q., Tamang E., Wintour A.. The Pareto Principle. The Plymouth Student Scientist, 7 (1):140-148, 2014.

[38] Hesselager, O. A Recursive Procedure for Calculations of Some Compound Distributions, Astin Bulletin, 24(1):19-32, 1994.

[39] Paul, S. R. Testing goodness of fit of the geometric distribution: an application to human fecundability data. Journal of Modern Applied Statistical Methods, 4:425–433, 2005.

[40] Ghosh, S., Resnick, S. A discussion on mean excess plots. Stochastic Processes and their Applications, 120(8):1492-1517, 2010.

[41] Nair, J, Wierman, A., Zwart, B. The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation. California Institute of Technology, 2020.

[42] Embrechts, P., Mikosch, T., Kluppelberg, C. Modelling Extremal Events: for Insurance and Finance. Springer-Verlag, London, UK, 1997.

[43] Dahab, A. Y., Hasbullah, H., Said, A. M. Predicting Traffic Bursts Using Extreme Value Theory. International Conference on Signal Acquisition and Processing, Kuala Lumpur, 2009.

[44] Neves, C., Alves, I. F. Ratio of Maximum to the Sum for Testing Super Heavy Tails. In: Advances in Mathematical and Statistical Modeling. Statistics for Industry and Technology. Birkhäuser Boston, USA, 2009.

[45] Bonetti, M., Cirillo, P., Musile Tanzi, P., Trinchero, E. An Analysis of the Number of Medical Malpractice Claims and Their Amounts. PLoS ONE 11(4): e0153362, 2016. doi:10.1371/journal.pone.0153362

[46] Arnold, T. B., Emerson, J. W. Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. The R Journal 3(2), 2011.

[47] Conover, W. J. A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions. Journal of American Statistical Association, Vol. 67, No. 339, 591–596, 1972.

[48] Gleser, L. J. Exact Power of Goodness-of-Fit Tests of Kolmogorov Type for Discontinuous Distributions. Journal of American Statistical Association, Vol. 80, No. 392, 954–958, 1985.

[49] Chen, J., Li, P. Hypothesis test for normal mixture models: The EM approach. Annals of Statistics, 37:2523–2542, 2009.

[50] Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M., Jordan, M., Yu, B. Sharp Analysis of Expectation-Maximization for Weakly Identifiable Models. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR 108:1866-1876, 2020.

[51] Queensley, C., Chukwudum, P. M., Mung'atu, J. K. Optimal threshold determination based on the mean excess plot, Communications in Statistics - Theory and Methods, In Press, 2019. https://doi.org/10.1080/03610926.2019.1624772

[52] Ghosh, S., Resnick, S. A discussion on mean excess plots. Stochastic Processes and their Applications, 120:1492–1517, 2010.

[53] Ghosh, S., Resnick, S. When Does the Mean Excess Plot Look Linear? Stochastic Models, 27(4), 2011.

[54] Feller, W. Introduction to Probability Theory and Its Applications, Wiley, 1971.

[55] Syakur, M. A, Khotimah, B. K., Rochman, E., Satoto, B. D. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. IOP Conference Series: Materials Science and Engineering, 2017.

[56] Albrecher, H., Beirlant, J., Teugels, J. Reinsurance: Actuarial and Statistical Aspects, Wiley, Chichester, 2017.

[57] Cirillo, P., Taleb, N. N. Tail risk of contagious diseases. Nature Physics, 16:606–613, 2020.

[58] Denic, S., Souid, A. K., Nicholls, M. G. The Automated Blood Count: Its History, Utility and Need for Change, 9(6), 2019.

[59] Chatterjee, P., Qi, M., Coe, N. B., Konetzka, R. T., Werner, R. M. Association Between High Discharge Rates of Vulnerable Patients and Skilled Nursing Facility Copayments. JAMA Internal Medicine, avalilable online, 2019. doi:10.1001/jamainternmed.2019.1209.

[60] Benford, F. The law of anomalous numbers. Proc. Am. Philos. Soc. 78(4): 551–572, 1938.

[61] Hellervik, A., Rodgers, G. A power law distribution in patients' lengths of stay in hospital, Physica A: Statistical Mechanics and its Applications 379(1), 2007.

[62] Gulati S., Shapiro, S. Goodness-of-Fit Tests for Pareto Distribution. In F Vonta (ed.), Statistical Models and Methods for Biomedical and Technical Systems, chapter 19, pp. 259-274. Birkhauser Basel, 2008.

[63] Munasinghe, R., Kossinna, P., Jayasinghe, D., Wijeratne, D. Package 'ptsuite'. https://cran.r-project.org/web/packages/ptsuite/

[64] Botimer, T. C. Efficiency considerations in airline pricing and yield management, Transportation Research Part A: Policy and Practice, 30(4): 307-317, 1996. https://doi.org/10.1016/0965-8564(95)00028-3.