A robustified modeling approach to analyze pediatric length of stay

ANDY H. LEE, PhD [1], MICHAEL GRACEY, MD PhD FAAP [1], KUI WANG, PhD [1] and KELVIN K.W. YAU, PhD [2]

[1] *Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, Perth, Western Australia*

[2] *Department of Management Sciences, City University of Hong Kong, Hong Kong*

Corresponding author:

Professor Andy H. Lee

Department of Epidemiology and Biostatistics

School of Public Health

Curtin University of Technology

GPO Box U 1987, Perth, WA 6845, Australia

Phone: +61-8-9266 4180

Fax: +61-8-9266 2958

E-mail: Andy.Lee@curtin.edu.au

Short running title: Analyzing pediatric length of stay

Number of tables = 2 **Number of figures** = 1

Word count of abstract = 181

Word count of main text = 2,166

ABSTRACT

PURPOSE: Length of stay (LOS) is an important measure of the cost of pediatric hospitalizations, but the guidelines developed so far are not rigorously evidence-based. This study demonstrates a robust gamma mixed regression approach to analyze the positively skewed LOS variable, which has implications for future studies of pediatric health care management.

METHODS: The robustified approach is applied to analyze hospital discharge data on childhood gastroenteritis in Western Australia (*n* = 514). The model accounts for demographic characteristics and co-morbidities of the patients, as well as the dependency of LOS outcomes nested within the 58 hospitals in the State. The method is compared to the standard linear mixed regression with trimming of extreme observations.

RESULTS: For the empirical application, the linear mixed regression results are sensitive to the magnitude of trimming. The identified significant factors from the robust regression model, namely infection, failure to thrive and iron deficiency anemia, are resistant to high-LOS outliers.

CONCLUSIONS: Robust gamma mixed regression appears to be a suitable alternative to analyze the clustered and positively skewed pediatric LOS, without transforming and trimming the data arbitrarily.

Key words: Gastroenteritis, Linear Mixed Model, Outliers, Robust Regression, Transformation

Abbreviations and acronyms

LOS = length of stay, ALOS = average length of stay, SD = standard deviation, *p* = p-value, CI = confidence interval

INTRODUCTION

Hospitalization of infants and children is an important component of the costs of providing health care funded by governments and other agencies. In developing countries and among other disadvantaged populations, childhood respiratory infections, diarrheal disease and malnutrition contribute to a high burden of illness and deaths [1]. Pediatric hospitalizations characteristically comprise a major budgetary item for governments, international health agencies and other health care providers. Therefore, comprehensive and accurate information about inpatient length of stay (LOS) should be a high priority for health planners and administrators in the strategic planning and deployment of financial, human and physical resources for these and related agencies [2-5]. In addition to resource allocation and quality control, there is also patient interest in anticipated dates of discharge [6]. However, pediatric LOS guidelines such as those developed by Milliman and Robertson are not rigorously evidence-based, but often drawn from benchmarking comparisons between institutions or group consensus rather than from epidemiologic data [7-9].

A comprehensive analysis of pediatric LOS has important implications in various aspects of health care management. In particular, the determination of patient-related characteristics affecting LOS can assist pediatricians to optimize care and rationalize their medical practice [3,4]. Moreover, performances of hospitals or plans, in terms of LOS, are not directly comparable without adjustment for patient casemix [10]. But the skewness of the LOS variable poses a problem for statistical modeling and analysis [7,11]. Current trimming methods to discriminate outliers from normal-stays are determined arbitrarily and without theoretical support [12]. In analyzing pediatric LOS, outliers, defined as the 2% of discharges with the longest LOS, are often excluded [7]. However, the choice is arbitrary and selection of an appropriate exclusion threshold can be difficult in small sample situations especially for complex diagnoses or conditions.

The objective of trimming in conjunction with a transformation (such as logarithmic) is to minimize the effects of extreme outliers and to attain the normality assumption on the LOS distribution, so that linear regression analysis can be applied. In this paper, a robustified hierarchical modeling approach is suggested as an alternative to linear mixed regression when analyzing heterogeneous LOS data. An empirical data set on pediatric LOS for infants admitted for gastroenteritis is used to illustrate the practical applications of the methodology.

METHODS

Linear Mixed Regression

When analyzing patient outcomes such as LOS, observations collected from the same hospital are often correlated. The dependence of clustered data (patients nested within hospitals) can lead to spurious associations and misleading inferences drawn from the standard regression model. A linear mixed regression model has been suggested to account for such dependency [10]. With the prediction of random hospital effects as a by-product of the model estimation process, hospital or plan performance can also be quantified and assessed. In view of the positive skewness of the empirical distribution, it has been suggested that the linear mixed regression model be applied to the logarithm of the trimmed LOS [10].

Let Y_{ij} be the logarithmic transformed pediatric LOS (number of days from admission to discharge) of the *j*th patient in the *i*th hospital; $i = 1,\ldots,m;\ j = 1,\ldots,n_i$ and $n = \sum_{i=1}^{m} n_i$ gives the total number of patients. The linear mixed regression model takes the form

$$Y_{ij} = X_{ij}\beta + u_i + e_{ij}$$

where X_{ij} represents the vector of covariates associated with the regression coefficients β (fixed effects), and the (random) hospital effect u_i and disturbances e_{ij} are assumed to have independent normal distributions with mean zero and variance σ_u^2 and σ_e^2, respectively. The method relates the mean LOS to the set of covariates but provides statistically efficient estimators and adjusted standard errors. The variance components reflect the respective variations at hospital and patient levels. The model estimation has been implemented in statistical packages such as SAS [13].

Robust Gamma Mixed Regression

Instead of linear mixed regression that relies on the normality assumption, a gamma mixed-effects regression model can be applied to analyze pediatric LOS and associated risk factors. The gamma distribution naturally accommodates different degrees of skewness of the underlying LOS distribution through its scale parameter. The gamma mixed regression model relates the mean LOS to the covariates and random hospital effects via a link function [14]. Model fitting and parameter estimation can be performed simply by invoking the SAS procedure GLIMMIX [13].

In the presence of extreme outliers, robust estimation for the gamma mixed regression model has been recommended to lessen their influence [15]. To achieve robustness in each iteration of the estimation procedure, whenever the standardized residual or hospital effect exceeds a prescribed boundary, its value will be replaced by the bound, thus limiting the effect of aberrant observations or hospitals on the regression coefficients. In statistical terminology, a bounded function is imposed to robustify the likelihood function so as to accomplish a desired degree of robustness in the model fitting. The most popular bounded influence function for robustification is the Huber ρ-function, whereas other bounded and redescending functions such as the Tukey biweight and Hampel piecewise linear functions can be used [15].

It should be noted that the robustified adaptation could render estimation of the gamma mixed regression model less statistically efficient. This is the tradeoff in avoiding potential spurious associations and misleading inferences induced by the LOS outliers. In general, a statistical efficiency of 90% can still be retained by adopting a Huber ρ-function with boundary value 2. Therefore, the boundary value of 2 is used in our empirical study. The robustified modeling approach appears to be a suitable alternative to analyze the clustered and skewed pediatric LOS without the need for arbitrary transformation and trimming of the data.

RESULTS

Gastroenteritis is an infectious disease prevalent among infants and children worldwide, especially in developing countries [1,16]. It is often associated with malnutrition and common in disadvantaged groups, particularly those living in overcrowded and unhygienic environments. The source of data for this empirical study is the Western Australia Hospital Morbidity Data System. Information on hospital discharge is routinely collected for all episodes of hospitalizations throughout the state of Western Australia. All infants born in 1995 hospitalized for gastroenteritis during their first year of life were included in our investigation, giving n = 514 patients clustered within m = 58 hospitals. The number of patients ranged from 3 to 190 per hospital for which linear mixed-effects and robust gamma mixed regressions are readily applicable. Gastroenteritis admissions were defined by *International Classification of Diseases, 9th Revision* clinical modification codes for principal diagnoses of diarrheal diseases. The outcome variable (pediatric LOS) is defined to be the number of days from admission to discharge. In addition, concomitant information on age at admission (months), sex, race, place of residence (urban/rural), admission type (emergency/elective), and co-morbidities namely the presence of dehydration, gastrointestinal sugar intolerance, failure to thrive, iron deficiency anemia and certain infections (genitourinary, scabies and/or otitis media), were extracted from the hospital separation records for each patient.

For this sample of infants admitted for gastroenteritis, the LOS ranged from 1 to 56 days and the median stay was 2 days. The empirical frequency distribution in Figure 1 shows that the LOS variable is positively skewed (skewness = 5.75). In order to minimize the effects of extreme outliers and to avoid analytical problems, the literature has suggested trimming LOS observations prior to statistical analysis [10]. The Western Australia health department currently uses a casemix funding formula for pediatric diagnoses whereby additional payments are made for high-LOS outliers 3 times the average LOS (ALOS). Consequently, the complete sample as well as the trimmed samples with thresholds set at 5, 7 and 10 days (3*ALOS) are considered, the corresponding patient characteristics and descriptive statistics are given in Table 1. For the complete sample, significant association was found between race and place of residence, as 80% of the indigenous patients resided in rural or remote areas of the State.

[Figure 1 and Table 1 about here]

Results from fitting the linear mixed-effects regression models to the logarithm of LOS are summarized in Table 2. It is evident that the subset of significant co-morbidities (dehydration, failure to thrive, iron deficiency anemia and infections) is sensitive to the magnitude of trimming used. The robust gamma mixed regression model with logarithmic link function is also fitted to the non-trimmed and untransformed LOS data. The logarithmic link function is adopted so that parameter estimates from the linear and gamma mixed regression models can be compared on the same scale. The robust fit generally agrees with the complete sample linear mixed regression with regard to the effects of the identified factors except dehydration. As expected, prolonged hospitalization for gastroenteritis is significantly associated with nutritional deficiencies and other illnesses such as infections, which may lead to complications and necessitate a longer period of treatment in hospital. The length of hospitalization also appears to be related to race. The group of 176 indigenous patients (ALOS=5.5, SD=7) stayed longer than the 328 non-indigenous patients (ALOS=2.3, SD=2.4); the differences in LOS being significant according to the nonparametric Mann-Whitney test ($p<0.001$). The delayed discharge of indigenous patients can often help to ensure drug compliance and to minimize the risk of re-infections, and partly because of logistical problems in arranging transport back to their remote home settlements. Dehydration upon presentation to hospital has some influence on LOS, but this co-morbidity only attains marginal significance according to the robust gamma regression fit ($p = 0.05$). For those variables associated with a prolonged stay, attention could be targeted at patients with the characteristics or at areas with large number of patients exhibiting such features.

[Table 2 about here]

The robustified modeling approach also provides information on differences in medical practice and discharge policy among hospitals. The random component σ_u estimate of 0.134 confirms hospital discrepancies as a relevant attribute of LOS variations. For linear mixed-effects regressions, however, it can be seen from Table 2 that this source of inter-hospital variations becomes negligible as the magnitude of trimming increases.

DISCUSSION

Gastroenteritis is prevalent among Australian Aboriginal infants and children and is a major cause of their hospitalization in Western Australia [17]. In the empirical study, infectious diarrheal diseases resulting in hospital admission were considered. Cases of mild gastroenteritis not requiring hospitalization were thus excluded. The proportion of episodes captured by the hospital admission data for the cohort was about 80% of the total number of notifications of enteric infections among infants less than one year old during the same period, based on data extracted from the Western Australian Notifiable Infection Disease Database (Communicable Disease Prevention and Control, Department of Health, Western Australia) [17]. We note that dehydration as co-morbidity was present in 27.8% of the infants at their initial admission for gastroenteritis. Dehydrated patients tended to stay on average 1.2 days more in hospital, but the evidence of the association between dehydration and LOS remains inconclusive. Oral rehydration therapy has been used successfully to treat mild to moderate diarrheal diseases in developing countries for more than thirty years and is now becoming more acceptable as appropriate therapy in the United States [18]. Children who present early for treatment for gastroenteritis and who are only mildly dehydrated, and do not have complicating co-existing illnesses, can often be rapidly rehydrated and managed in a non-hospital environment and later supervised in less expensive and more child-friendly settings [19]. However, data comparing the differences in treatment between those diagnosed as dehydrated and those without dehydration were not available for further assessment.

The identification and quantification of risk factors affecting pediatric LOS, after adjusting for random hospital effects, are important for clinical practice, discharge planning, and health care management of patients. By targeting relevant factors influencing LOS, appropriate policies can be developed to manage the hospital care and the health care resources effectively, as well as the early prediction of infants requiring a longer period of hospitalization. The information obtained also assists in the planning of hospital bed requirements. In addition, early identification of children at high risk of prolonged LOS will allow physicians to treat them more aggressively, and permit their families to estimate the costs of hospital care and the anticipated day of discharge.

The robust gamma mixed regression advocated here represents an evidence-based and more practical way of analyzing LOS than the standard linear mixed regression approach. The advantage is that transformation of LOS and trim points for long-stay outliers need not be defined a priori and arbitrarily. Instead, all observations contribute to the estimation of model parameters. Gamma regression is also more appropriate statistically than linear regression for small or moderate sample sizes when the normality assumption cannot be satisfied.

Although our empirical application focuses on gastroenteritis hospitalizations, the robustified modeling approach is applicable to analyze other pediatric LOS. Moreover, instead of hospital as the unit of clustering, the methodology can be adapted to accommodate variations in other settings. For example, patients may be nested under different health plans or local districts within the state. Such a hierarchical analysis is of interest to state health authorities to assess variations in service consumption between different population subgroups. Similarly, the robustified modeling approach can be used to analyze longitudinal data, where repeated episodes of some events such as recurrent gastroenteritis for each patient are monitored.

Acknowledgements

This research was supported by grants from the National Health and Medical Research Council of Australia (project #229003) and the Research Grants Council of Hong Kong. We would like to thank Diana Rosman, William Pyper, Peter Somerford of the Department of Health, Western Australia, for providing the hospitalization data and coding information. A computer program to fit the robust gamma mixed regression model is available upon request to the corresponding author.

REFERENCES

1. UNICEF. *The State of the World's Children 2003*. New York, NY; 2002.
2. Malkin JD, Keeler E, Broder MS, Garber S. Postpartum length of stay and newborn health: A cost-effectiveness analysis. *Pediatrics* 2003; 111: e316-322.
3. Bianco A, Pileggi C, Trani F, Angelillo IF. Appropriateness of admissions and days of stay in pediatric wards of Italy. *Pediatrics* 2003; 112: 124-128.
4. Srivastava R, Homer CJ. Length of stay for common pediatric conditions: teaching versus nonteaching hospitals. *Pediatrics* 2003; 112: 278-281.
5. Silber JH, Rosenbaum PR, Even-Shoshan O, Shabbout M, Zhang X, Bradlow ET, Marsh RR. Length of stay, conditional length of stay, and prolonged stay in pediatric asthma. *Health Serv Res* 2003; 38: 867-886.
6. Zernikow B, Holtmannspötter K, Michel E, Hornschuh F, Groote K, Hennecke KH. Predicting length-of-stay in preterm neonates. *Eur J Pediat*. 1999; 158: 59-62.
7. Sills MR, Huang ZJ, Shao C, Guagliardo MF, Chamberlain JM, Joseph JG. Pediatric Milliman and Robertson length-of-stay criteria: Are they realistic? *Pediatrics* 2000; 105: 733-737.
8. Bauchner H, Vinci R, Chessare J. Milliman and Robertson-going in the wrong direction. *Pediatrics* 2000; 105: 858-859.
9. Harman JS, Kelleher KJ. Pediatric length of stay guidelines and routine practice: the case of Milliman and Robertson. *Arch Pediatr Adolesc Med* 2001; 155: 885-890.
10. Leung KM, Elashoff RM, Rees KS, Hasan MM, Legorreta AP. Hospital- and patient-related characteristics determining maternity length of stay: A hierarchical linear model approach. *Am J Public Health*. 1998; 88: 377-381.
11. Marazzi A, Paccaud F, Ruffieux C, Beguin C. Fitting the distributions of length of stay by parametric models. *Med Care* 1998; 36: 915-927.

12. Lee AH, Xiao J, Vemuri SR, Zhao Y. A discordancy test approach to identify outliers of length of hospital stay. *Statist Med* 1998; 17: 2199-2206.

13. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc; 1996.

14. McGilchrist CA. Estimation in generalized mixed models. *J Roy Statist Soc Ser B* 1994; 56: 61-69.

15. Yau KKW, Kuk AYC. Robust estimation in generalized linear mixed models. *J Roy Statist Soc Ser B* 2002; 64: 101-117.

16. Yousafzai MA, Bhutta ZA. Global epidemiology of childhood diarrhoea at the turn of the millennium. In: Bhutta ZA, Ed. *Contemporary issues in Childhood Diarrhoea and Malnutrition.* Oxford: Oxford University Press; 2000: 1-22.

17. Lee AH, Flexman J, Wang K, Yau KKW. Recurrent gastroenteritis among infants in Western Australia: A seven-year hospital-based cohort study. *Ann Epidemiol* 2004; 14: 137-142.

18. Santosham M. Oral Rehydration Therapy. *Arch Pediatr Adolesc Med* 2002; 156: 1177-1179.

19. McConnochie KM, Conners GP, Lu E, Wilson C. How commonly are children hospitalized for dehydration eligible for care in alternative settings? *Arch Pediatr Adolesc Med* 1999; 153: 1233-1241.

TABLE 1. Patient demographic and descriptive statistics

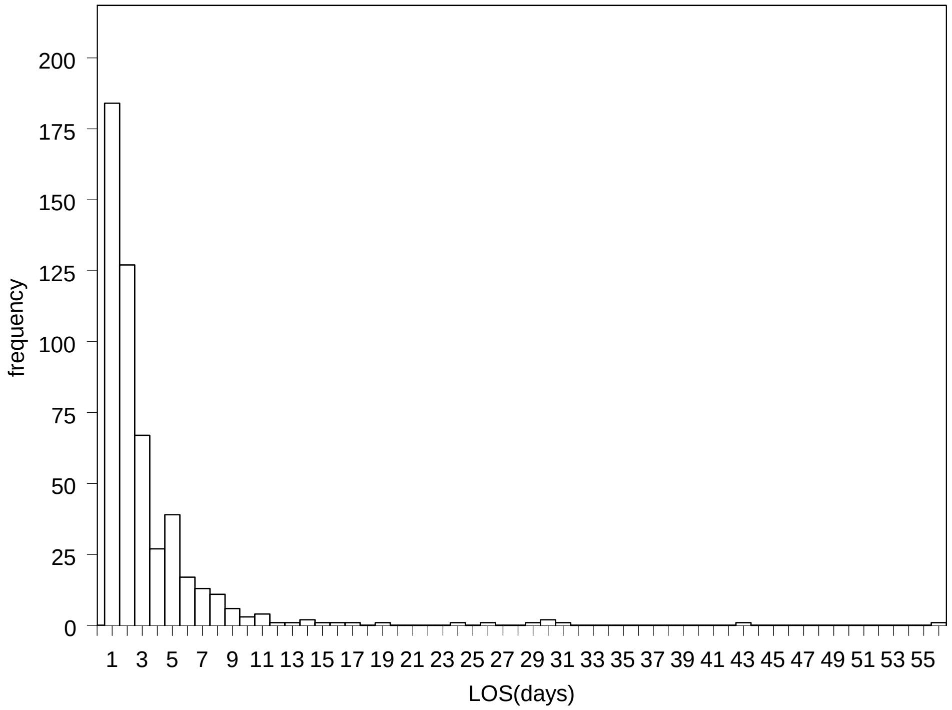
	Complete sample	Trimmed sample (LOS ≤ 10 days)	Trimmed sample (LOS ≤ 7 days)	Trimmed sample (LOS ≤ 5 days)
Number of patients *n*	514	494	474	444
Number of hospitals *m*	58	58	58	57
Skewness	5.75	1.48	1.20	1.01
ALOS in days	3.39	2.65	2.39	2.12
(SD)	(4.78)	(2.01)	(1.61)	(1.26)
Mean admission age in months (SD)	6.96 (3.19)	7.01 (3.20)	7.07 (3.19)	7.07 (3.18)
Proportion of patients (%)				
Male	54.9	55.1	54.6	55.2
Indigenous patient	34.2	32.2	30.8	27.7
Rural residence	50.8	50.6	50.4	48.9
Elective admission	9.5	9.7	9.7	10.4
Dehydration	27.8	26.5	26.2	25.2
Gastrointestinal sugar intolerance	8.0	6.9	6.1	4.7
Failure to thrive	4.9	4.1	3.4	2.3
Iron deficiency anemia	3.3	2.6	1.7	1.1
Infection (genitourinary/scabies/otitis media)	13.4	12.2	11.2	10.6

TABLE 2. Coefficients and associated 95% confidence intervals of factors affecting gastroenteritis LOS from fitting robust gamma mixed regression and linear mixed effects regression models

	Robust gamma mixed regression	Mixed regression[a]	Mixed regression[a] (≤ 10 days)	Mixed regression[a] (≤ 7 days)	Mixed regression[a] (≤ 5 days)
n	514	514	494	474	444
Intercept	0.558[c] (0.327, 0.789)	0.429[c] (0.213, 0.645)	0.464[c] (0.274, 0.654)	0.447[c] (0.271, 0.623)	0.462[c] (0.315, 0.609)
Admission age in months	-0.003 (-0.023, 0.017)	-0.008 (-0.026, 0.010)	-0.001 (-0.019, 0.017)	0.005 (-0.011, 0.021)	0.005 (-0.011, 0.021)
Male	-0.053 (-0.176, 0.070)	-0.025 (-0.139, 0.089)	-0.034 (-0.140, 0.072)	-0.066 (-0.168, 0.036)	-0.053 (-0.153, 0.047)
Indigenous	0.587[c] (0.428, 0.746)	0.530[c] (0.385, 0.675)	0.452[c] (0.319, 0.585)	0.431[c] (0.304, 0.558)	0.357[c] (0.232, 0.482)
Rural residence	-0.008 (-0.194, 0.178)	0.027 (-0.149, 0.203)	-0.012 (-0.165, 0.141)	-0.013 (-0.154, 0.128)	-0.057 (-0.167, 0.053)
Elective admission	0.041 (-0.182, 0.264)	0.064 (-0.142, 0.270)	0.013 (-0.173, 0.199)	-0.045 (-0.225, 0.135)	-0.019 (-0.184, 0.146)
Dehydration	0.146 (-0.003, 0.289)	0.153[b] (0.024, 0.282)	0.089 (-0.033, 0.211)	0.087 (-0.031, 0.205)	0.066 (-0.052, 0.184)
Gastrointestinal sugar intolerance	0.745[c] (0.510, 0.980)	0.700[c] (0.486, 0.914)	0.619[c] (0.409, 0.829)	0.549[c] (0.333, 0.765)	0.427[c] (0.190, 0.664)
Failure to thrive	0.594[c] (0.294, 0.894)	0.531[c] (0.257, 0.805)	0.416[c] (0.142, 0.690)	0.360[b] (0.072, 0.648)	0.134 (-0.207, 0.475)
Iron deficiency anemia	0.530[c] (0.162, 0.898)	0.572[c] (0.237, 0.907)	0.527[c] (0.184, 0.870)	0.380 (-0.026, 0.786)	0.282 (-0.206, 0.770)
Infection	0.263[c] (0.071, 0.455)	0.249[c] (0.075, 0.423)	0.152 (-0.015, 0.319)	0.116 (-0.049, 0.281)	0.107 (-0.060, 0.274)
Random hospital component σ_u	0.134	0.140	0.100	0.075	0.001

[a] outcome variable is log(LOS) [b] $p < 0.05$ [c] $p < 0.01$

FIGURE 1. Empirical frequency distribution of gastroenteritis LOS (*n* = 514)


frequency
LOS(days)