Predicting the Whole Distribution with Methods for Depth Data Analysis Demonstrated on a Colorectal Cancer Treatment Study

D. Vicendese[1,2(✉)], L. Te Marvelde[1,2], P. D. McNair[2], K. Whitfield[2], D. R. English[1,3], S. Ben Taieb[4], R. J. Hyndman[4], and R. Thomas[5]

[1] Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, Melbourne, VIC, Australia
D.Vicendese@latrobe.edu.au
[2] Cancer Strategy and Development, Department of Health and Human Services, Melbourne, VIC, Australia
[3] Centre for Epidemiology and Biostatistics, The University of Melbourne, Carlton, VIC, Australia
[4] Department of Econometrics and Business Statistics, Monash University, Clayton, VIC, Australia
[5] Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, VIC, Australia

Abstract. We demonstrate the utility of predicting the whole distribution of an outcome rather than a marginal change. We overcome inconsistent data modelling techniques in a real world problem. A model based on additive quantile regression and boosting was used to predict the whole distribution of length of hospital stay (LOS) following colorectal cancer surgery. The model also assessed the association of hospital and patient characteristics over the whole distribution of LOS. The model recovered the empirical LOS distribution. A counterfactual simulation quantified change in LOS over the whole distribution if an important associated predictor were to be varied. The model showed that important hospital and patient characteristics were differentially associated across the distribution of LOS. Model insights were much richer than just focusing on a marginal change. This method is novel for public health and epidemiological studies and could be applied in other fields of research.

Keywords: Additive quantile regression · Machine learning · Boosting · Density forecast

1 Introduction

Evidence of an association between hospital or surgeon cancer surgery volume and better patient outcomes has been mixed [1–7], however, previous analyses have had important limitations. A 2002 review that examined seven statistical modelling techniques used to assess association between patient factors and length of hospital stay (LOS) in a cardiovascular setting found that choice of model influenced the conclusion

H. Nguyen (Ed.): RSSDS 2019, CCIS 1150, pp. 162–182, 2019.
https://doi.org/10.1007/978-981-15-1960-4_12

of the analyses. In some instances the conclusions were reversed. Model results were inconsistent due to LOS being a complex phenomenon and unmet assumptions regarding distributional fit. A small proportion of patients with very long hospital stays made it inherently difficult for simpler parametric methods to effectively model the data [8]. These models have been employed also in colorectal cancer (CRC) studies of association between LOS and provider volume and hence some doubt is cast on both negative and positive conclusions [1, 9–12]. Some studies have used arbitrary thresholds for the categorization of LOS and volumes (low, medium and high) [1, 2, 11, 13, 14] which may reduce statistical power [15]. If the skewness of LOS changes as provider volume changes then arbitrary categorizations of LOS may be problematic [15–17]. Additionally, heterogeneity due to arbitrary categorizations has prevented synthesis of findings [3, 18].

A weakness of some hospital patient outcome studies is bias due to unaccounted for correlations between outcomes within a hospital - sometimes referred to as random effects [19]. Furthermore, surgeries are performed in the context of a hospital with a distinct infrastructure and management that affect their outcome [2]. Modelling this contextual effect may yield important information regarding its association with patient outcomes such as LOS [20–22]. The systematic differences in patients' outcomes across hospitals that persist after differences in patients' risk profiles have been accounted for reflect differences in hospitals' quality of care [23]. In this study we examined the relationship between LOS following CRC surgery and provider volume by using a quantile regression model which makes no distributional assumption about LOS or error terms [24, 25] and avoids arbitrarily categorizing LOS or provider volume [15]. The model formulation we specified took the individual patient as the unit of analysis and used the clustering of patients within a hospital to analyse the association between the hospital context and LOS [20–22]. Furthermore, we did not focus on just a marginal change, such as a mean or median or some other quantile, but instead modelled the whole distribution so as to give a more in depth understanding of the interplay between LOS and hospital and patient characteristics over the whole distribution of LOS [26].

2 Methods

2.1 Data Details

The Victorian Admitted Episode Dataset (VAED) includes all separations (discharges and transfers) undertaken within all Victorian hospitals. All separations between 1 Jul 2005 and 30 Jun 2015 recorded in the VAED, that included one of 30 ICD-10-AM Australian Classification of Health Interventions procedure codes for colorectal surgery, as the primary reason for admission, were identified. There were 62,774 admissions for 57,446 patients. Analysis was restricted to admissions whose principal diagnosis was for CRC, ICD-10-AM codes C18.x to C21.x which resulted in a final data set of 28,343 admissions for 27,633 patients. Provider volume was defined as the number of colorectal surgical procedures performed by a hospital within a fiscal year (1 July to 30 June), whether patients had a principal diagnosis of CRC or not. That is,

annual volume (AV). Length of stay was defined as the number of days from admission to discharge for the episode of care including transfers to other hospitals and geriatric and rehabilitation centres.

2.2 Modelling Details

As it was conceivable that LOS and provider volume were not necessarily linearly related, we used an additive quantile regression (AQR) model which does not require a predetermined functional fit but instead determines the best fit from the data [15, 25, 27, 28]. The specification we used was based on a formulation by Mundlak [29] and is in a class commonly referred to as a 'within and between' effects model [20, 22, 30]. It required that we enter both AV and mean annual volume (MAV). Mean annual volume (MAV) was defined for each hospital as the mean of all AV over the number of years the hospital operated within the 10 year study period. Not all hospitals performed colorectal surgical procedures in every study year [2]. The within effect was modelled by AV. It estimated the effect on LOS within hospitals as AV varied and its interpretation is equivalent to any fixed effect estimator [30]. The between effect was modelled by MAV. Due to the model formulation used, it estimated the effect on LOS if a patient were to attend another hospital with a different MAV, that is, the hospital contextual effect [20, 21, 23]. This method draws comparisons across hospitals and estimates the effect of hospital choice on patient LOS, or in other words, hospitals' quality of care or efficiency regarding LOS [23]. The model was adjusted for various patient and hospital factors that may confound the association between provider volume and LOS [19, 31–33].

We tested how well the model represented the data by predicting the empirical cumulative distribution (CDF) of LOS and assessed its fit. We refer to this as the recovered distribution. As the model formulation estimated the effect on LOS if a patient were to attend another hospital [20, 21, 23], we used the model to simulate the change in LOS if CRC patients were to have counterfactually attended a hospital that the model indicated to be more LOS efficient. We predicted each percentile 1 to 99 which were combined to obtain the predicted CDF of LOS [26]. This was termed the counterfactual CDF. The area under the counterfactual CDF was calculated and compared to the area under the recovered CDF. As the area under each CDF directly related to the total sum of LOS days, the difference in areas estimated the change in total sum of LOS due to this hypothetical experiment. This enabled us to calculate a dollar value for the difference by allowing $1000 per LOS day [34]. A 95% confidence interval (CI) was computed for the estimated change. Statistical significance was set at the 0.05 level.

2.2.1 Boosting to Assess Variable Selection and Functional Fit

To aid model building and variable selection we used boosting [15, 26, 35–37]. Boosting is a statistical algorithm in the class of machine learning methods. It determines the most appropriate model by optimizing the fit to the data while limiting over fitting. It will discard variables, along with their proposed functional fit that do not aid optimization where optimization is defined as the largest reduction in the model fit error at each iterative step. For the actual form of the loss function to determine model fit

error, please see [38]. Hence it is a variable and functional form selection method that does not resort to heuristic techniques such as ad hoc stepwise variable selection [24]. It also works well in the setting of many predictors with possibly high correlation between them [38]. We used component-wise gradient boosting [38]. Continuous variables may be entered simultaneously as linear and non-linear components into the model and the boosting process is able to determine which variable and which functional fit aids model fit to the signal in the data. Categorical variables are entered linearly. This process is thoroughly described in the following references [15, 24, 26, 37]. Intrinsic to this method is the choice of step length and optimal number of iterations. We used the default of 0.1 for step length. General cross validation (GCV) was used to choose 5000 as an optimal number of iterations. We carried this out with the freeware R version 3.3.3 [39], using the package mboost [24, 40]. All continuous variables were mean centered as recommended for the boosting algorithm [24, 25]. We assessed possible random effects [14] by including a random intercept for hospital into the boosting process where hospital was represented by a dummy variable.

2.2.2 Model Specification for Additive Quantile Regression with Boosting

The following model was implemented using the boosting algorithm:

$$
\begin{aligned}
&Q_t(\text{LOS}) \text{ modelled with } \textit{intercept} \\
&+ f_{nl}(\text{mean annual surgery volume}) + f_l(\text{mean annual surgery volume}) \\
&+ f_{nl}(\text{annual surgery volume}) + f_l(\text{annual surgery volume}) \\
&+ f_{nl}(\text{age}) + f_l(\text{age}) \\
&+ f_{nl}(\text{daily surgery admissions}) + f_l(\text{daily surgery admissions}) \\
&+ f_{nl}(\text{Elixhauser comorbidity score}) + f_l(\text{Elixhauser comorbidity score}) \\
&+ f_{nl}(\text{year of discharge}) + f_l(\text{year of discharge}) \\
&+ f_{nl\,cyclic}(\text{month of discharge}) + f_l(\text{month of discharge}) \\
&+ f_l(\text{sex}) + f_l(\text{ASA}) + f_l(\text{cancer site}) + f_l(\text{metastatic cancer}) \\
&+ f_l(\text{laparoscope}) + f_l(\text{admission type}) + f_l(\text{separation mode}) \\
&+ f_l(\text{hospital type}) + f_l(\text{collocated}) + + f_l(\text{surgical procedure}) \\
&+ f_{ri}(\text{hospital campus}),
\end{aligned}
\tag{1}
$$

where: Q_t = modelled quantile, t = 1 to 99;

f_{nl} = non linear function;

f_l = ordinary linear squares function for continuous or categorical data;

$f_{nl\,cyclic}$ = non linear function using cyclic splines;

f_{ri} = random intercept to adjust for correlations in LOS outcomes within a hospital.

2.2.3 Recovering the Unconditional Predicted Quantile

As with ordinary linear regression that predicts a mean of a dependent variable conditional on independent covariates, quantile regression also predicts quantiles of the distribution conditional on independent covariates. That is, if we refer to the quantiles

we are modelling as $Q(z)$ for the z^{th} quantile and the vector of covariates as $\boldsymbol{x}$, then we are modelling $\widehat{Q}(z|\boldsymbol{x})$, where $\widehat{Q}$ indicates an estimate of Q. To obtain the unconditional estimate $Q(z)$ we need to average out the effect of $\boldsymbol{x}$ over the distribution that the sample is drawn from. That is,

$$Q(z) = \int_{-\infty}^{+\infty} f(\boldsymbol{x})\hat{Q}(z|\boldsymbol{x})dx \tag{2}$$

where $f(\boldsymbol{x})$ is the joint probability distribution of the covariates. We don't know the mathematical formulation for $f(\boldsymbol{x})$, however, the above expression is equivalent to $E\left[\widehat{Q}(z|\boldsymbol{x})\right]$, that is, the mean of $\widehat{Q}(z|\boldsymbol{x})$. Since our sample is clearly representative of $f(\boldsymbol{x})$ and the number of estimates is very large (28,343 = the sample size), taking the mean will converge to $Q(z)$ by the law of large numbers.

2.2.4 Smoothing Count Data – a Technicality

Due to the combination of LOS being a count variable and the estimation of the conditional quantile involves a non-smooth objective function, a certain amount of smoothness needs to be imposed on the data. This is done by adding a specific form of random noise, referred to as jittering, which preserves the one to one relationship between the jittered and un-jittered data. Hence the model estimates based on the jittered data are readily converted to their un-jittered values. This is well described by Machado and Silva [41]. To accomplish this, we used the dither function in the R package quantreg [42]. The counts were recovered from the modelled jittered LOS by applying Theorem 2 by Machado and Silva [41],

$$\widehat{Q}(z|\boldsymbol{x}) = \text{ceiling}\left[\text{jittered}\left\{\widehat{Q}(z|\boldsymbol{x})\right\} - 1\right] \tag{3}$$

where $\widehat{Q}(z|\boldsymbol{x})$ **is** the estimated quantile conditional on $\boldsymbol{x}$, the vector of covariates, and ceiling[n] denotes the smallest integer greater than or equal to n.

2.2.5 Building the Second Additive Quantile Regression Model Without Boosting

We used the results of the boosting to build a second AQR model. The boosting indicated that random effects for hospital was not important for predicting LOS and that all entered variables may be important for predicting LOS except for Elixhauser (co-morbidity) score, co-location status and sex – see Fig. 1. Month and separation mode were the strongest predictors in the sense of reduction of model fit error by up to approximately 3 and 1 respectively for some of the percentiles. Other variable contributions were below about 0.5 for all percentiles. Although boosting suggested that sex was not an important variable, we still included it due to its generally important relevance in epidemiological studies. This decision was eventually justified, see Sect. 3.3.4 below. Boosting also suggested that non-linear fits for MAV, AV, age and month better predicted LOS than linear but a linear fit for year and number of daily colorectal surgery admissions better predicted LOS. We still entered the number of

daily admissions and year as a non-linear fit as model regularization would likely result in a linear fit without taxing degrees of freedom.

The variables that were indicated as important for predicting LOS were then entered into the second AQR. This was done without including random effects for hospital as suggested by the boosting. The AQR was applied with a regularization of the model fit to the data to reduce over fitting and to increase prediction accuracy. This is done by the calculation of smoothing parameters for the continuous variables and the use of a least absolute shrinkage and selection operator (lasso) for the categorical variables [27, 35, 36]. The selection of the smoothing parameter for each continuous variable is first initiated on a univariate basis. These initial values are then passed onto a function, along with a starting value for the lasso, which then recalibrates them over the whole parameter space for each model fit to each percentile. These functions and the R computer code for their implementation are well described by Koenker [27].

This is the specification of the second AQR model that we implemented following variable selection by the boosting process.

$$\begin{aligned} &Q_t(\text{LOS}) \text{ modelled with } \textit{intercept} \\ &+ f_{nl}(\text{mean annual surgery volume}) + f_{nl}(\text{annual surgery volume}) \\ &+ f_{nl}(\text{age}) + f_{nl}(\text{daily surgery admissions}) + f_{nl}(\text{year of discharge}) \\ &+ f_l(\text{month of discharge}) + f_l(\text{sex}) + f_l(\text{ASA}) + f_l(\text{cancer site}) \\ &+ f_l(\text{metastatic cancer}) + f_l(\text{laparoscope}) + f_l(\text{admission type}) \\ &+ f_l(\text{separation mode}) + f_l(\text{hospital type}) + f_l(\text{surgical procedure}). \end{aligned} \tag{4}$$

The second AQR was carried out with the R package quantreg [42] and was compared to the boosted model.

2.2.6 Comparing the AQR Models - with Boosting to Without Boosting

The models were compared in four ways. Firstly we used the continuous ranked probability score (CRPS) [26, 43, 44] as formulated by Gneiting and Ranjan to compare density forecasts [43] - see Eq. 6 in their paper. Secondly, we calculated the Akaike information criterion (AIC) for each percentile and added the results. We defined the AIC as 2 * (number of variables - logarithm of the likelihood). Thirdly we compared by eye how well the graph of the recovered distribution represented the actual empirical distribution of LOS. We termed this graphical fit. Fourthly, we computed the areas under the graphs of the empirical and recovered distribution, between the 1st and 99th percentiles and compared them. A recovered distribution should give an area close to the area under the empirical distribution. Computing the area under the graph was done using the R package flux using the auc function [45].

Table 1 compares model capability of both models in recovering the empirical distribution of LOS. Although the boosted model had overall lower AIC, due to the importance of the continuous ranked probability score (CRPS) for comparing density forecasts, the superior approximation of the area under the graph and better graphical fit (Fig. 2) we proceeded with the second AQR model in assessing the association between provider volume and LOS.

Table 1. Model performance indicators. See Sect. 2.2.6.

	Boosted model	2nd AQR
CRPS Lower is better	4.27	3.95
Total AIC over 1–99% Lower is better	12,173,948	19,758,132
Area under graph ratio to empirical LOS Close to 1 is better	0.95	1.02

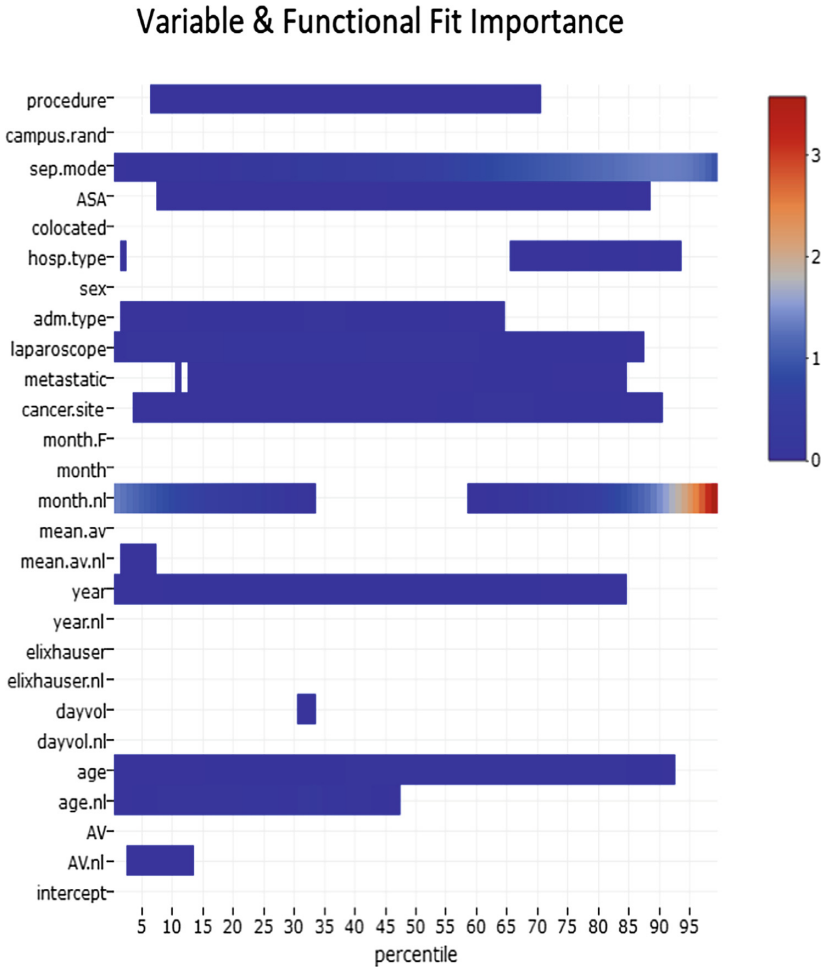

Fig. 1. Variable and functional fit importance indicated by boosting. The scale is total reduction in model fit error over all iterations. See Sects. 2.2.1 and 2.2.5.

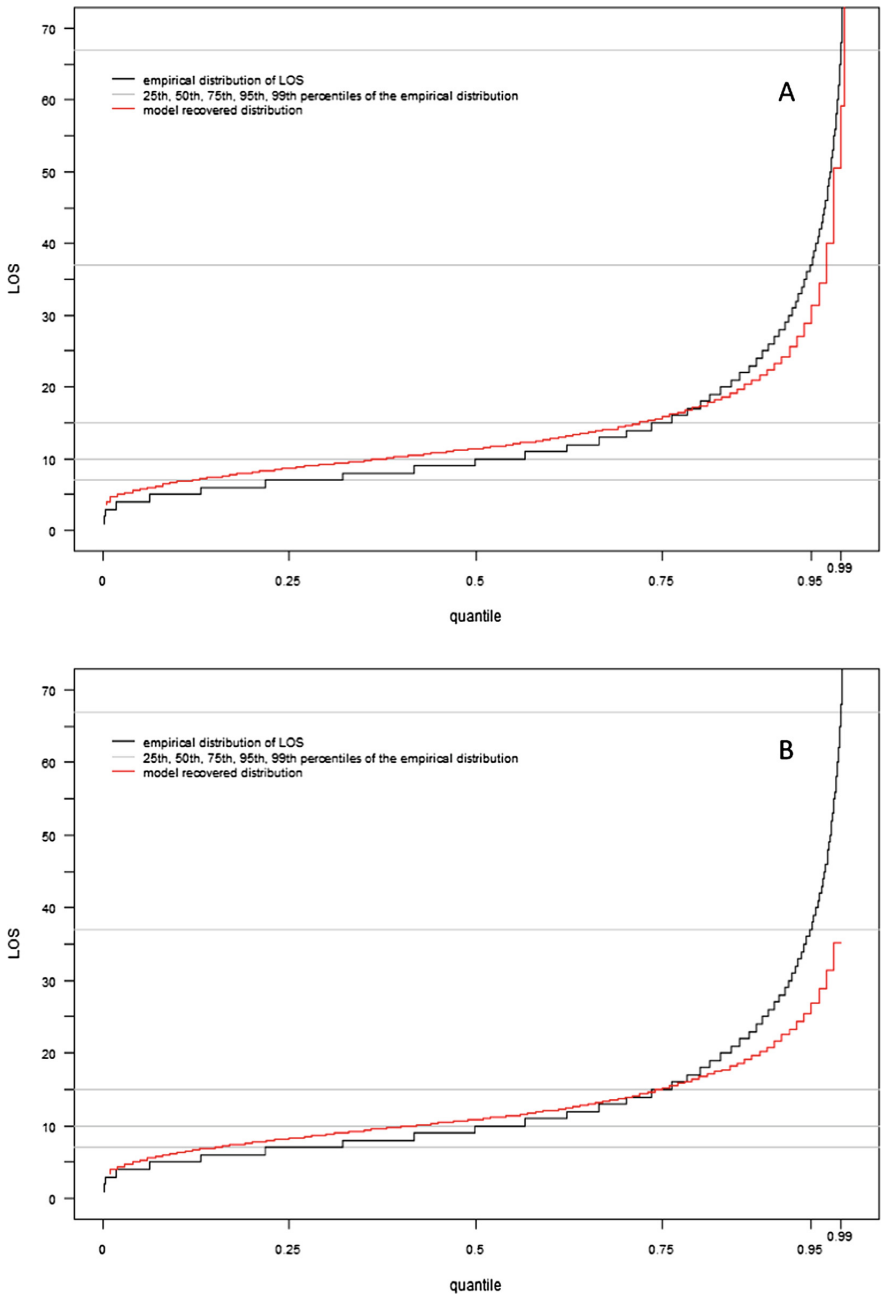

Fig. 2. The empirical distribution of LOS in black compared to the recovered distribution by the AQR models in red: **A** without boosting; **B**: with boosting. (Color figure online)

3 Results

For the sake of brevity we present the important results that illustrate the utility of our model.

3.1 Mean Annual Volume Association with LOS

The model graphs (Fig. 3) indicate that hospitals' performances (contextual effect) regarding patient LOS varied greatly for all percentiles and that this variation was not systematically associated with MAV. At almost any point in the graph, looking left or right displays both lower and higher LOS. The graphs display initial falls in LOS for MAV up to 33 approximately which are then followed by marked variation, with low points in LOS at MAV of 122.1 and 245.8 and a high point at 105.6 MAV. The former two MAV had the lowest LOS over all percentiles 40 and 20 times respectively while the latter had the highest 89 times. For all percentiles, the p values from an F statistic for model fit were less than 1×10^{-6}.

3.2 Counterfactual Prediction of Change in LOS Contingent on Change in MAV

To carry out the counterfactual prediction, we selected the hospital with MAV of 122.1 as an LOS efficient hospital due to its consistent association with reduced LOS over many percentiles. There were 68 hospitals that had MAV of 122.1 or less and they generated 106,488 LOS days (27.5%) from 7,979 episodes of care (28.1%). The counterfactual prediction estimated a fall of 8.5% in total LOS over all patients and hospitals, $p < 0.007$, with 95% CI (2.9%, 14.6%). This equated to a reduction of 32,842 total LOS days, 95% CI (11,225, 56,434 days) or a predicted saving of about $32 million in present day terms over the ten year study period by allowing approximately $1000 per LOS day [34]. Figure 4 demonstrates that, for the counterfactual experiment, the predicted savings mainly came from reduced LOS for patients who had LOS between percentiles 14 and 94.

By using the model's counterfactual prediction capacity, an index of performance with a 95% CI can be obtained for any hospital. The prediction would immediately indicate if the hospital was performing better, worse or at about the same level as all other hospitals. If performing the counterfactual experiment, using the hospital being assessed as the basis of comparison, resulted in a 5% drop in total LOS then that hospital would have an index of .95 with an associated confidence interval. Lower than 1 indicates superior efficiency, higher than 1 inferior and 1 no difference. This index is independent of any distributional or model fit assumptions or arbitrary categorization. For the analysis of LOS, the Victorian Auditor General's report resorted to using trimmed data in a linear regression. This method may be subject to statistical objections if a mean is not representative of the whole data and because information in the tail(s) of the distribution, that may represent patients who may not fit the profile of a mean patient, is discarded [25, 34, 46, 47].

Our model could be used in national or international settings as it can allow for nesting in those levels, and so help assess hospital efficiency in regard to LOS in broader contexts. This would assist with synthesis of future international studies when in the past, diverse categorization methods had impeded synthesis. It can be extended to analyse variation between hospitals regarding other outcomes such as mortality and readmission following CRC surgery.

3.3 Further Results for Patient and Hospital Factors

3.3.1 Laparoscope Use – See Fig. 5

Use of laparoscope was associated with reduced LOS. There were growing reductions of between 0.5 to 4 days with increasing LOS percentile which indicated the importance of laparoscope use, where possible, for helping to reduce unnecessarily long LOS. Use of laparoscope was the main modifiable feature of our analysis. This suggested where possible, use of laparoscope is important for patient outcomes and resource allocation, as observed and recommended by others [11]. This is an example of where, if solely a marginal change had been predicted, the result would have been just one coefficient comparing laparoscope use to non – use. The one coefficient would have represented an overall average effect and would have been graphically illustrated as a flat line across all percentiles. Instead, with quantile regression, we see how the effect due to laparoscope use varied over across all percentiles of LOS. This observation holds for all the presented variable results. All p values for all coefficients over all percentiles were less than 1×10^{-4}.

3.3.2 Separation Mode – See Fig. 6

It is evident from Fig. 6 that discharge to transition care, an aged care residential facility, other acute hospital or a statistical separation contributed heavily to protracted LOS with between 2–50 days greater than patients who were transferred to home. This difference increased with increasing LOS percentile. Left against medical advice shows no association with LOS as the coefficients for nearly all percentiles are close to zero and non-significant. It seems that patients who died in hospital after colorectal cancer surgery did so either early in their hospital stay or at the end of a protracted stay. The p values for all statistically significant coefficients, were mainly less than 0.01.

There may be some scope for improvements in LOS for patients who are transferred for extended care to other acute hospitals, rehabilitation or geriatric care centres. These patients had between 2–50 days more LOS compared to patients who were transferred to home which accounted for 104,497 days (27%) of total LOS. If with vigilant follow up and management, a modest 10% of these days were to be saved, that would have amounted to a saving of about 10,400 days and a potential saving of about $10 million in present terms over the 10 year study period [34].

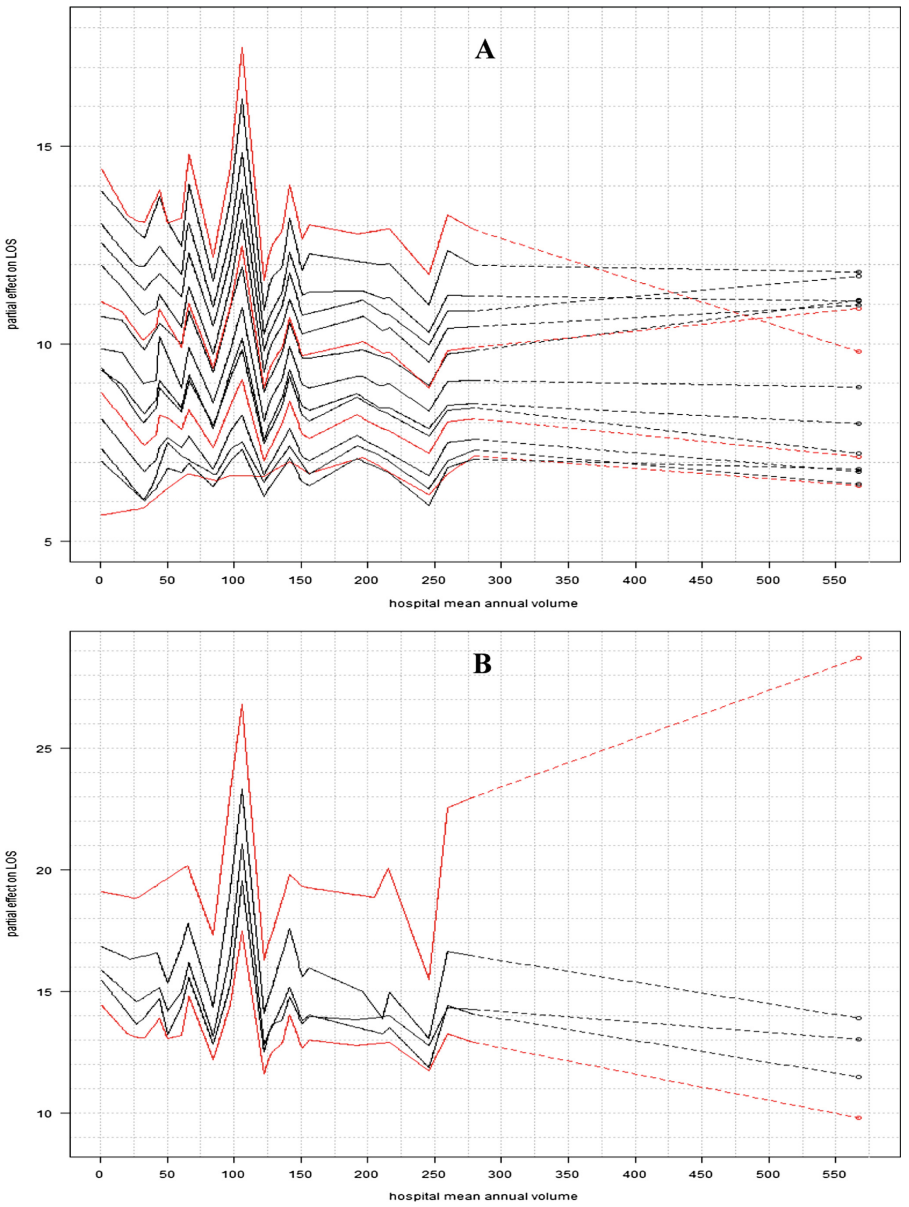

Fig. 3. Model estimates for between hospital differences (contextual effect) in the association between annual volumes and LOS; **A**-percentiles 5–75, **B-**percentiles 75–95, both in intervals of 5. Red lines are percentiles 5, 25, 50, 75 and 95. The lines are dotted between 279.6 and 566.7 as there were no hospitals with MAV between these values. A few of the percentile fits display quantile crossing (see Sect. 3.4). All p values, from an F statistic that assessed model fit for percentiles 1–95, were less than 1×10^{-6}. (Color figure online)

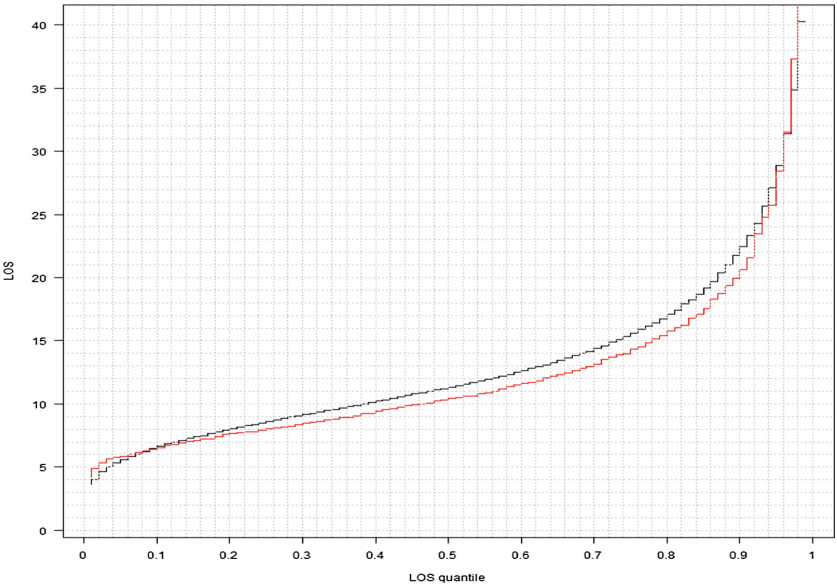

Fig. 4. In black we have the recovered distribution. Red is the counterfactual distribution obtained by setting all MAV to 122.1 and all AV to the same AV in each year as generated by the hospital with MAV of 122.1. (Color figure online)

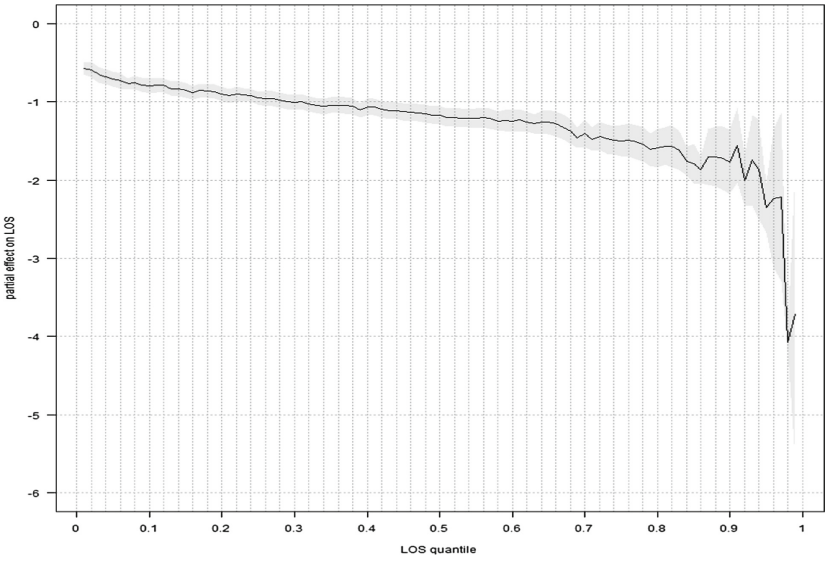

Fig. 5. The association between laparoscope use and LOS. Use of laparoscope is compared to non-use of laparoscope. The shading represents the coefficient 95% CI for each percentile.

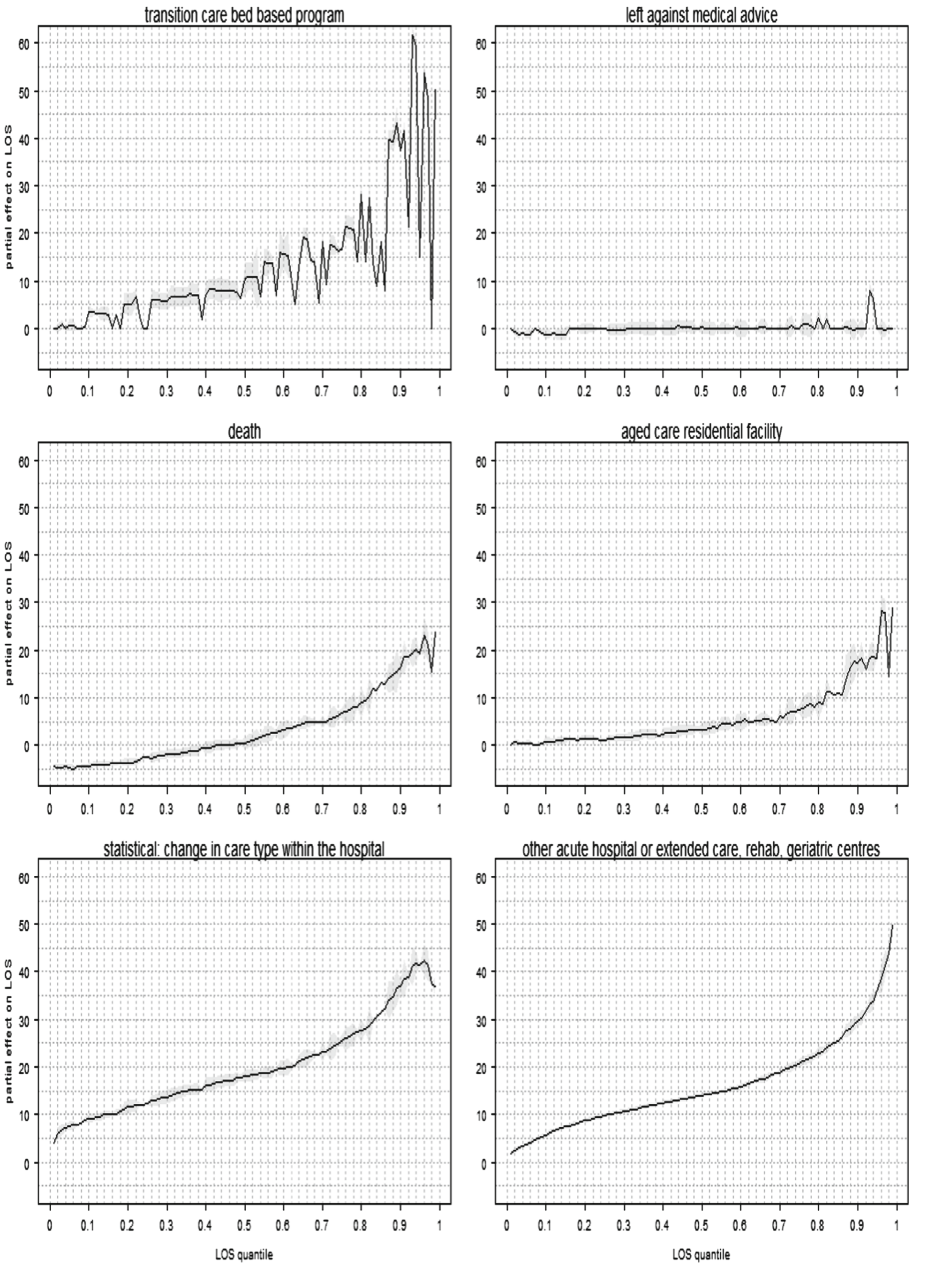

Fig. 6. Model estimates for the association between separation mode and LOS for all percentiles 1–99. These coefficients use separation to private residence or accommodation as the reference level. The shading represents the coefficient 95% CI for each percentile.

3.3.3 Month of Year – See Fig. 7

Please note the months are within a fiscal year and are ordered from July to June. Compared to July, month of year differences in LOS ranged between −4.6 to 3.9 days over all percentiles but the bulk of differences were between −0.5 to 0.5 days. The larger magnitudes were for the higher LOS percentiles of 90 or more. Except for January, all months mostly had lower LOS than July but only May, June, October and December showed consistency (Fig. 7C) and strong statistical evidence (Fig. 7B) for this difference. January tended to have higher LOS however, zero was not extreme for this relative difference compared to July (Fig. 7C) as can also be seen by lack of statistical significance (Fig. 7B). Most likely these monthly effects were due to seasonally adjusted hospital administration factors rather than an environmental seasonal effect.

The association between month of year and LOS was a surprise finding. This is more likely to reflect seasonally modified discharge practices rather than seasonal environmental factors [48] and so may potentially be another modifiable feature. If the same efficiencies that seem to have applied to discharges in May, June, December and October were to be applied to discharges in all other months of the year, then potentially there could have been reductions of approximately 1000–2000 days of LOS which would translate to savings of between $1 million to $2 million dollars approximately over the 10 year study period.

Adding these savings and savings due to better management of separation mode (see Sect. 3.3.2) to savings reaped from increased hospital efficiency indicated in the main results (see Sect. 3.2), potential total savings due to improved LOS efficiency could be more than $4 million per year for CRC surgery.

3.3.4 Sex - See Fig. 8

For percentiles up to about 20 there were no statistical differences between LOS for women and men. For higher percentiles we see a gradual decrease for women to about 3 days less LOS compared to men. This may be related to being in better general health at surgery due to health promoting behaviours that are more likely to be exhibited by women than men.

3.4 Quantile Crossing

When the predicted unconditional quantiles are combined to recover the full distribution, the monotonicity of the cumulative distribution function (CDF) may not be retained. That is, at times a predicted value of LOS that does not respect the strictly increasing property of quantiles may be produced by a model. This has been referred to as quantile crossing [25, 28]. We used the process of monotone rearranging to restore the required monotonicity property to the CDF for the recovered distribution of LOS [26, 49]. The rearrangement process was based on mathematical results by Hardy, Littlewood and Polya [50, 51] and was implemented with the rearrangement function in the R package, quantreg [42].

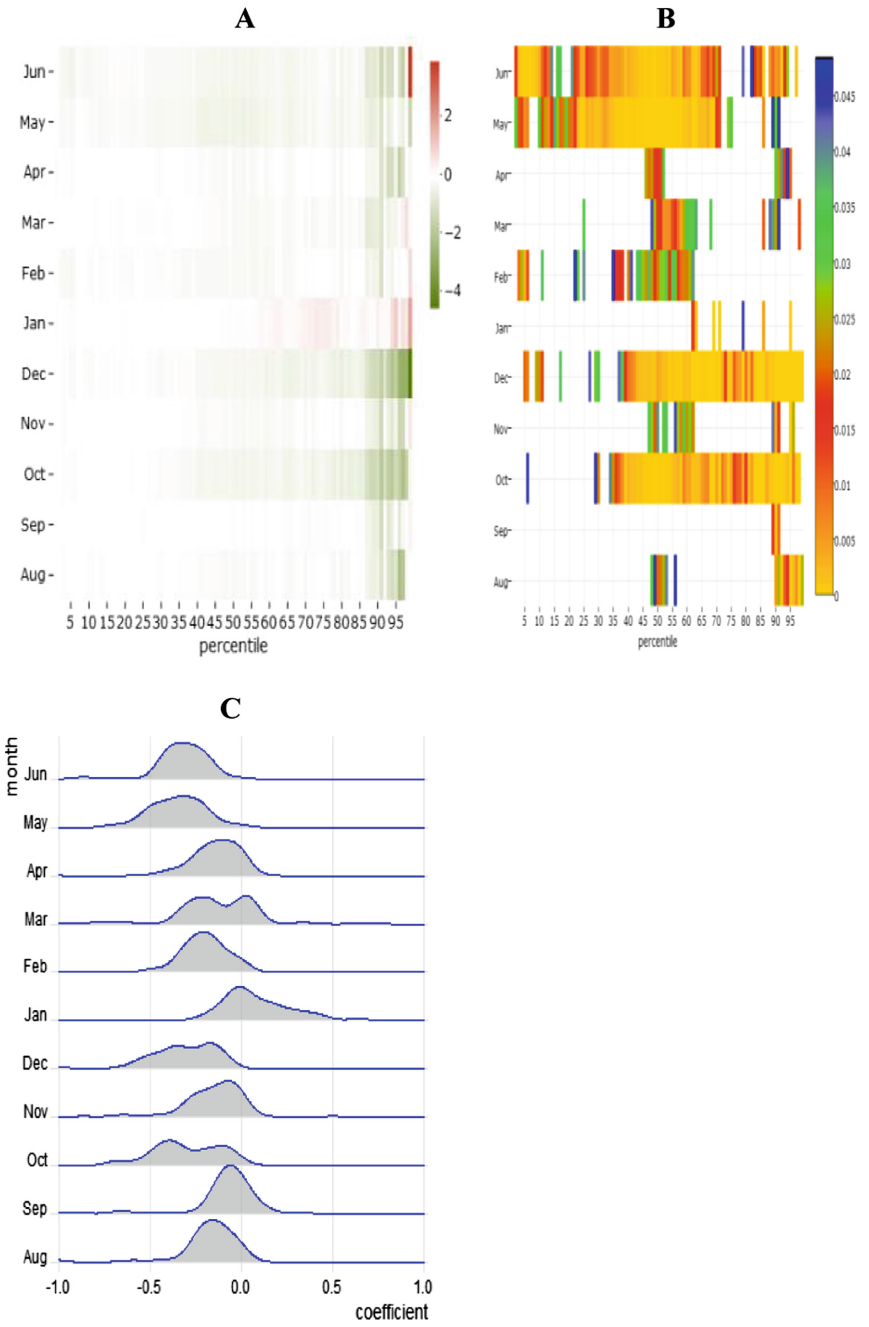

Fig. 7. **A** - Coefficients for month of year compared to July for percentiles 1–99. **B -** p values for month coefficients for percentiles 1–99. White indicates a p value > 0.05. **C -** Distribution of coefficients for each month.

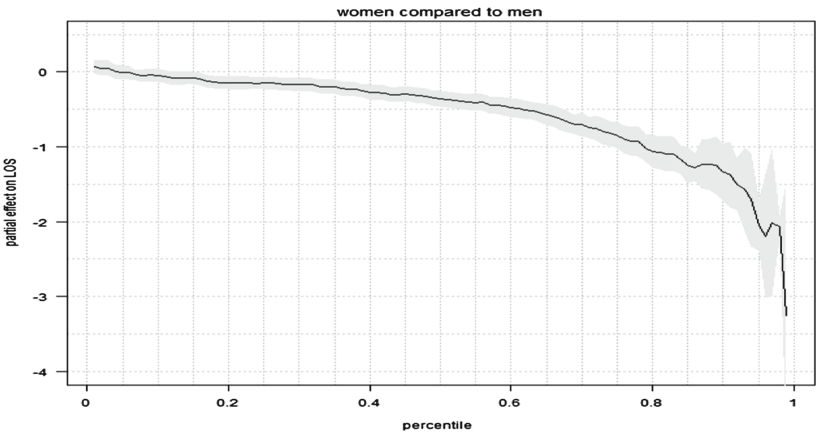

Fig. 8. Model estimates for the association between sex and LOS for all percentiles 1–99. These coefficients are based on comparison of women to men as the reference level. The shading represents the coefficient 95% CI for each percentile.

Quantile crossing may also occur when generally modelling the association between different quantiles of a dependent variable as a *function* of an independent variable. This occurred in this study a small number of times in modelling the association between quantiles of LOS and mean annual volume, annual volume and age of patient – only the results for mean annual volume, Fig. 3, are displayed for brevity purposes. It is evident that fitted lines (functions) for a few of the percentiles cross the lines of other percentiles. The advantage of quantile regression in not requiring global distributional assumptions but instead using the data near the specified quantile, sometimes has the disadvantage of producing quantile crossing. That is, in modelling one quantile, quantile regression is ignorant of other quantiles [25]. However, quantile crossing usually occurs in sparse areas of the data and is also sensitive to outlier values for the *independent* covariate (quantile regression is robust to *dependent* variable outlier values) [25, 52]. Linear regression may also be affected by these conditions [25]. The 99^{th} LOS percentile for age was affected by both conditions as the 99^{th} percentile is a sparse region of the data and there was a patient of 103 years of age. For mean annual volume and annual volume, there was a clear outlier with a relatively large distance from the next lowest values and no intervening values (dotted lines in those graphs). The fit determined by data to the left of the dotted line cannot anticipate lack of data to the right. If quantile crossing occurs substantially often, then this may indicate model misspecification [25]. Quantile crossing is an area of ongoing research [25, 49].

4 Discussion

Our model presented itself as a useful method for assessing provider performance regarding LOS, adjusted for pertinent patient and hospital factors [19]. The additive quantile regression approach proved a suitable method to cope with the difficulties

inherent in analysing LOS due to its greatly skewed distribution and non-linear association with provider volume. Assuming a distributional fit or arbitrary categorization of LOS and focusing only on the mean would not have allowed the assessment of associations with LOS over its entire distribution. Imposing a linear fit causes non-linear associations to go undetected, if present in the data. Focusing on only one parameter, such as a mean, would have assumed that change in mean LOS would have translated equally to all patients or that the "average" patient was representative of all patients. It is quite possible that there may be important changes in the tails of a distribution yet, the mean is unchanged [25]. A linear, Cox or logistic regression would have resulted in only one (average) coefficient for all percentiles and would not have given any insight into differential associations across LOS percentiles. That is, focusing only on mean outcomes may cause important information about patients who may not fit well into the frame of average to go unobserved.

General linear models focus on marginal change in the outcome variable dependent on predictive factors, however it is quite possible that there is a multifaceted interchange between the outcome and predictive factors which may vary for different levels of the outcome. If we aim to understand the full impact of possible predictors on an outcome, predicting a marginal change is not sufficiently meaningful [26, 53–55]. To model a possible multifaceted interchange between predictors and outcomes, we need to examine change in the outcome, conditional on the predictor, across the outcome distribution in its entirety or at least a sufficiently representative number of percentiles. Our model obtained a more realistic idea of the change in the outcome variable, conditioned on important covariates, by predicting all percentiles. This has been likened to the difference between calculating only "what changes" compared to "who changes" and by how much [46] where, in our case, surgery patients were the who. These are ongoing and important issues for epidemiological studies and data analysis generally [47].

Our model did not assume any distributional fit which may bias estimates or reduce sensitivity to detect associations if the assumption is an over simplification and therefore not justified [56]. John Tukey referred to this as having more honest foundations for data analysis [47]. Furthermore, we did not use any arbitrary categorical definitions of the outcome, or continuous predictors which has hindered synthesis of past studies. John Ioannidis in his seminal work, Why Most Published Research Papers are False, stated in his 4th Corollary that the greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true [57].

Although we used a Mundlak formulation in our example, it is not necessary to do so. The basic idea of predicting the outcome distribution can be used in any typical regression formulation for a continuous or count outcome. Our model was further complemented by shrinkage and penalization to obtain more accurate estimates and reduce statistical error [19, 27].

Our model did have the limitation of being data hungry. A large sample size is required to reliably employ our model especially if many covariates are entered into the analysis. However, this age of big data somewhat mitigates this limitation. Our model is also computer resource intensive. To predict percentiles 1–99 and run the counterfactual simulation took approximately 70 min on a 64-bit PC with 64 GB RAM and a

Xeon® 3.5 GHz CPU but 5.5 h on a 64-bit PC with 16 GB of RAM and an i5 2.5 GHz CPU. The recalibration of the lasso and smoothing parameters over the whole parameter space for each model fit for percentiles 1–99, took approximately 36 h on a 64-bit PC with 16 GB of RAM and an i5 2.5 GHz CPU.

This model can only be applied to continuous or count outcomes but there has been some recent work that has applied quantile regression to time to event data and work that seeks to apply it to binary response data [25].

5 Conclusion

Predicting the whole outcome distribution was useful in providing an in depth description of the complexity of the associations between hospital and patient factors across the whole distribution of LOS. The facility of the model to indicate change over the whole distribution is useful where predicting a change in mean or median outcome is an oversimplification of the data and does not provide insight that is sufficiently indicative and real-world. With sufficient data, our model can be applied to any continuous or count outcome. Improvements in optimizing the time to predict the percentiles and calculate lasso and smoothing parameters are required. As far as we can tell, this method is novel for public health and epidemiological studies and may have further uses in these areas as well as other fields of scientific research.

References

1. Borowski, D.W., Bradburn, D.M., Mills, S.J., Bharathan, B., Wilson, R.G., Ratcliffe, A.A., et al.: Volume-outcome analysis of colorectal cancer-related outcomes. Br. J. Surg. **97**(9), 1416–1430 (2010)
2. Burns, E.M., Bottle, A., Almoudaris, A.M., Mamidanna, R., Aylin, P., Darzi, A., et al.: Hierarchical multilevel analysis of increased caseload volume and postoperative outcome after elective colorectal surgery. Br. J. Surg. **100**(11), 1531–1538 (2013)
3. Chowdhury, M.M., Dagash, H., Pierro, A.: A systematic review of the impact of volume of surgery and specialization on patient outcome. Br. J. Surg. **94**(2), 145–161 (2007)
4. Faiz, O.: The volume–outcome relationship in colorectal surgery. Tech. Coloproctol. **18**(10), 961–962 (2014). Official Journal of SICCR, MSCP, ISCRS, ECTA, Colorectal Anal Group of Surgical Section of Chinese Medical Association, MSPFD
5. Killeen, S.D., O'Sullivan, M.J., Coffey, J.C., Kirwan, W.O., Redmond, H.P.: Provider volume and outcomes for oncological procedures. Br. J. Surg. **92**(4), 389–402 (2005)
6. Kizer, K.W.: The volume-outcome conundrum. New Engl. J. Med. **349**(22), 2159–2161 (2003)
7. McGrath, D.R., Leong, D.C., Gibberd, R., Armstrong, B., Spigelman, A.D.: Surgeon and hospital volume and the management of colorectal cancer patients in Australia. ANZ J. Surg. **75**(10), 901–910 (2005)
8. Austin, P.C., Rothwell, D.M., Tu, J.V.: A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. Health Serv. Outcomes Res. Method. **3**(2), 107–133 (2002)

9. Gatt, M., Anderson, A.D., Reddy, B.S., Hayward-Sampson, P., Tring, I.C., MacFie, J.: Randomized clinical trial of multimodal optimization of surgical care in patients undergoing major colonic resection. Br. J. Surg. **92**(11), 1354–1362 (2005)
10. Huebner, M., Hubner, M., Cima, R.R., Larson, D.W.: Timing of complications and length of stay after rectal cancer surgery. J. Am. Coll. Surg. **218**(5), 914–919 (2014)
11. Thompson, B.S., Coory, M.D., Gordon, L.G., Lumley, J.W.: Cost savings for elective laparoscopic resection compared with open resection for colorectal cancer in a region of high uptake. Surg. Endosc. **28**(5), 1515–1521 (2014)
12. Zheng, Z., Hanna, N., Onukwugha, E., Bikov, K.A., Mullins, C.D.: Hospital center effect for laparoscopic colectomy among elderly stage I-III colon cancer patients. Ann. Surg. **259**(5), 924–929 (2014)
13. Faiz, O., Haji, A., Burns, E., Bottle, A., Kennedy, R., Aylin, P.: Hospital stay amongst patients undergoing major elective colorectal surgery: predicting prolonged stay and readmissions in NHS hospitals. Colorectal Dis.: Off. J. Assoc. Coloproctol. Great Br. Irel. **13** (7), 816–822 (2011)
14. Gruen, R.L., Pitt, V., Green, S., Parkhill, A., Campbell, D., Jolley, D.: The effect of provider case volume on cancer mortality: systematic review and meta-analysis. CA: Cancer J. Clin. **59**(3), 192–211 (2009)
15. Fenske, N., Fahrmeir, L., Hothorn, T., Rzehak, P., Hohle, M.: Boosting structured additive quantile regression for longitudinal childhood obesity data. Int. J. Biostat. **9**(1), 1–8 (2013)
16. Borghi, E., de Onis, M., Garza, C., Van den Broeck, J., Frongillo, E.A., Grummer-Strawn, L., et al.: Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. Stat. Med. **25**(2), 247–265 (2006)
17. Wang, X., Dey, D.K.: Generalized extreme value regression for binary response data: an application to B2B electronic payments system adoption. Ann. Appl. Stat. **4**(4), 2000–2023 (2010)
18. Archampong, D., Borowski, D., Willejrgensen, P., Iversen, L.H.: Workload and surgeons specialty for outcome after colorectal cancer surgery. Cochrane Colorectal Cancer Group **3** (3) (2012)
19. Ash, A.S., Fienberg, S.F., Louis, T.A., Norm, S.T., Stukel, T.A., Utts, J.: Statistical issues in assessing hospital performance. Committee of Presidents of Statistical Societies The COPSS-CMS White Paper Committee (2011). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.352.6798
20. Bell, A., Fairbrother, M., Jones, K.: Fixed and random effects models: making an informed choice 2017 March 2018. https://www.researchgate.net/publication/299604336_Fixed_and_Random_effects_models_making_an_informed_choice
21. Feaster, D., Brincks, A., Robbins, M., Szapocznik, J.: Multilevel models to identify contextual effects on individual group member outcomes: a family example (Report). Fam. Process **50**(2), 167 (2011)
22. Dieleman, J.L., Templin, T.: Random-effects, fixed-effects and the within-between specification for clustered data in observational health studies: a simulation study. PLoS ONE **9**(10), e110257 (2014)
23. Danks, L., Duckett, S.: All complications should count: using our data to make hospitals safer (Methodological supplement) (2018). https://grattan.edu.au/wp-content/uploads/2018/02/897-All-complications-should-count-methodological-supplement.pdf
24. Hofner, B., Mayr, A., Robinzonov, N., Schmid, M.: Model-based boosting in R: a hands-on tutorial using the R package mboost. Comput. Stat. **29**(1), 3–35 (2014)
25. Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge (2005)

26. Taieb, S.B., Huser, R., Hyndman, R.J., Genton, M.G.: Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. IEEE Trans. Smart Grid **7**(5), 2448–2455 (2016)
27. Koenker, R.: Additive models for quantile regression: model selection and confidence bandaids. Braz. J. Probab. Stat. **25**(3), 239–262 (2011)
28. Koenker, R.: Quantile Regression in R: A Vignette, 6 March 2018 (2018). https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf
29. Mundlak, Y.: On the pooling of time series and cross section data. Econometrica **46**(1), 69 (1978)
30. van de Pol, M., Wright, J.: A simple method for distinguishing within - versus between-subject effects using mixed models. Anim. Behav. **77**(3), 753–758 (2009)
31. Aravani, A., Samy, E.F., Thomas, J.D., Quirke, P., Morris, E.J., Finan, P.J.: A retrospective observational study of length of stay in hospital after colorectal cancer surgery in England (1998–2010). Medicine **95**(47), e5064 (2016)
32. Cologne, K.G., Byers, S., Rosen, D.R., Hwang, G.S., Ortega, A.E., Ault, G.T., et al.: Factors associated with a short (<2 Days) or Long (>10 Days) length of stay after colectomy: a multivariate analysis of over 400 patients. Am. Surg. **82**(10), 960–963 (2016)
33. Field, K., Shapiro, J., Wong, H.L., Tacey, M., Nott, L., Tran, B., et al.: Treatment and outcomes of metastatic colorectal cancer in Australia: defining differences between public and private practice. Intern. Med. J. **45**(3), 267–274 (2015)
34. Frost, P.: Victorian auditor-general's report: hospital performance: length of stay. In: Victorian, Auditor-General's, Office (eds.) (2016)
35. Efron, B., Hastie, T.: Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press, Cambridge (2016) https://web.stanford.edu/∼hastie/CASI_files/PDF/casi.pdf
36. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009) https://web.stanford.edu/∼hastie/Papers/ESLII.pdf. https://doi.org/10.1007/978-0-387-84858-7
37. Hofner, B., Hothorn, T., Kneib, T., Schmid, M.: A framework for unbiased model selection based on boosting. J. Comput. Graph. Stat. **20**(4), 956–971 (2011)
38. Bühlmann, P., Hothorn, T.: Boosting algorithms: regularization, prediction and model fitting. Stat. Sci. **22**(4), 477–505 (2007)
39. R Core Team: R: a language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing (2017) http://www.R-project.org/
40. Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., Hofner, B.: mboost: Model-Based Boosting (2017) https://CRAN.R-project.org/package=mboost
41. Machado, J.A.F., Silva, J.M.C.S.: Quantiles for counts. J. Am. Stat. Assoc. **100**(472), 1226–1237 (2005)
42. Koenker, R.: Quantreg: Quantile Regression (2017) https://CRAN.R-project.org/package=quantreg
43. Gneiting, T., Ranjan, R.: comparing density forecasts using threshold - and quantile-weighted scoring rules. J. Bus. Econ. Stat. **29**(3), 411–422 (2011)
44. Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. **15**(5) (2000) https://journals.ametsoc.org/action/doSearch?AllField=Decomposition+of+the+Continuous+Ranked+Probability+Score+for+Ensemble+Prediction+Systems
45. Jurasinski, G., Koebsch, F., Guenther, A., Beetz, S.: Flux: flux rate calculation from dynamic closed chamber measurements (2014). https://CRAN.R-project.org/package=flux

46. Hohl, K.: Beyond the average case: the mean focus fallacy of standard linear regression and the use of quantile regression for the social sciences. SSRN, Elsevier (2009). https://ssrn.com/abstract=1434418
47. Tukey, J.: More honest foundations for data analysis. J. Stat. Plan. Inference **57**(1), 21–28 (1997)
48. Kc, D.S., Terwiesch, C.: Impact of workload on service time and patient safety: an econometric analysis of hospital operations. Manag. Sci. **55**(9), 1486–1498 (2009)
49. Chernozhukov, V., Fernández-Val, I., Galichon, A.: Quantile and probability curves without crossing. Econometrica **78**(3), 1093–1125 (2010)
50. Burchard, A.: A short course on rearrangement inequalities (2018). https://www.math.toronto.edu/almut/rearrange.pdf
51. Hardy, H.G., Littlewood, J.E., Pólya, G.: Inequalities. Acta Applicandae Mathematica **23**(1), 95 (1991)
52. John, O.O.: Robustness of quantile regression to outliers. Am. J. Appl. Math. Stat. **3**(2), 86–88 (2015)
53. Kneib, T.: Beyond mean regression. Stat. Model. **13**(4), 275–303 (2013)
54. Harvey, A.: Discussion of 'Beyond mean regression'. Stat. Model. **13**(4), 363–372 (2013)
55. Le Cook, B., Manning, W.G.: Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. Shanghai Arch. Psychiatry **25**(1), 55–59 (2013)
56. Fenske, N., Burns, J., Hothorn, T., Rehfuess, E.A.: Understanding child stunting in India: a comprehensive analysis of socio-economic, nutritional and environmental determinants using additive quantile regression. PLoS ONE **8**(11), e78692 (2013)
57. Ioannidis, J.P.: Why most published research findings are false. PLoS Med. **2**(8), e124 (2005)