

# Predicting surgical case durations using ill-conditioned CPT code matrix

Ying Li<sup>\*</sup>, Saijuan Zhang<sup>†</sup>, Reginald F. Baugh<sup>‡</sup> and Jianhua Z. Huang<sup>†</sup>

## Abstract

Efficient utilization of existing resources is crucial for cost containment in medical institutions. Accurately predicting surgery duration will improve the utilization of indispensable surgical resources such as surgeons, nurses, and operating rooms. Prior research has identified the current procedural terminology (CPT) codes as the most important factor when predicting surgical case durations. Yet there has been little reporting of a general predictive methodology that can effectively extract information from multiple CPT codes. In this research, we propose two regression-based predictive models, a linear regression and a log-linear regression. To perform these regression analysis, a full-ranked design matrix based on CPT code inclusions in the surgical cases needs to be constructed. However, naively constructed design matrix is ill-conditioned (i.e. singular). We devise a systematic procedure to construct a full-ranked design matrix by sifting out the CPT codes without any predictive power while retaining useful information as much as possible. Our proposed models can be applied in general situations where a surgery can have any number of CPT codes and any combination of CPT codes. Using real world surgical data, we compare the proposed models with benchmark methods and find remarkable reductions in prediction errors.

KEY WORDS: healthcare, data analysis, regression analysis

---

<sup>\*</sup>Corresponding author, Department of Information and Operations Management, Mays Business School, Texas A&M University, College Station, TX 77843. Telephone: 979-458-4787. Fax: 979-845-5653. Email: yli@mays.tamu.edu.

<sup>†</sup>Department of Statistics, Texas A&M University, College Station, TX 77843

<sup>‡</sup>Division of Otolaryngology, The University of Toledo Medical Center, Toledo, OH 43614

# 1 Introduction

The pressure of cost containment makes efficient utilization of existing resources a top priority for medical institutions. Surgeons, nurses and operating rooms are indispensable for a surgery to be performed, and are important resources for a medical service provider. The schedule of these resources should be based upon surgical case durations (Weiss 1990, Olivares et al. 2008), which, however, can be highly variable. Such variability poses a serious challenge to surgical scheduling and resource utilization (Litvak and Long 2000, McManus et al. 2003). Accurately predicting surgical case duration is a pressing need in hospital management.

Surgery, by nature, involves a series of physical activities. Each surgery is characterized by one or multiple current procedural terminology (CPT) codes. A CPT code is a five-digit number that represents a set of medical, surgical or diagnostic services. The CPT code or the combination of CPT codes that is prescribed for a surgery dictates the core actions taken during the surgery. CPT codes are maintained by the American Medical Association for uniformity. Naturally, CPT codes are a key factor that determines the duration of a surgery. Having surveyed the articles in the area of general thoracic surgery, Dexter et al. (2008) find that the precise procedure types, which are represented by CPT codes, were the *most important* factor when predicting surgical case durations. Ignoring the critical information conveyed in CPT codes will often lead to unsatisfactory predictions of surgical case durations. For instance, Combes et al. (2008) use data mining tools to predict the duration of surgeries. Their results, while shedding lights on the benefit of applying data warehousing models, are reportedly not satisfactory. The authors believe that their grouping of surgeries based on diagnoses rather than procedure types is the main reason for inaccuracy. Motivated by the need to incorporate surgery type information, we present in this paper predictive methods for surgical case durations based directly on the CPT codes included in each surgery.

Modeling surgical case durations has been a topic of interest for operations management and medical communities. Given the role CPT codes play in a surgery, the majority of the existing literature involve CPT codes directly or indirectly. There are two major lines of approaches among existent work utilizing CPT related information. One relies on linear regressions for estimating surgical case durations or identifying the crucial factors that affect variability in surgeries. In a multi-phase study, Wright et al. (1996) find that surgeons provide better time estimates than the scheduling software adopted in their institution. Based on this finding, Wright et al. (1996) develop regression models for predicting surgical case durations by including, as the explanatory variables (or independent variables), the surgeons' own estimates, the estimates from the scheduling soft-

ware, and several other characteristics of a surgery. CPT codes are not directly included as part of the explanatory variables in their regression models. But surgeons are aware of the CPT codes prescribed for a surgery when making their estimates. As such, CPT codes are implicitly utilized. The regression models are shown to outperform both surgeons and the scheduling software. This study supports the inclusion of CPT codes as the explanatory variables for predicting surgical case durations. Strum et al. (2000) investigate factors associated with variability in surgery durations. They select surgeries with a single CPT code that were repeatedly performed by one or multiple surgeons. A five-factor main-effects linear model is established for each CPT code under consideration. This study identified surgeon as the second most important source of variability after the CPT codes.

The other line of work studies the fitness of known distributions, notably the normal distribution and the lognormal distribution, for the purpose of predicting surgery case durations. Strum et al. (2000) examine a large set of real surgery data and test how well the lognormal and normal distributions fit the data set. In their study, only surgeries with a single CPT code are considered, and the surgeries are categorized based on its CPT-anesthesia combination. Goodness-of-fit tests are conducted for each of those CPT-anesthesia combinations. Strum et al. (2000) conclude that lognormal distributions fit the surgery data better than normal distributions. They also note that the Shapiro-Wilk goodness-of-fit test can sometimes reject lognormal distributions that seemingly fit the surgery data, and thereby suggest using normal probability plots together with goodness-of-fit tests. The lognormal distributions investigated by Strum et al. (2000) have two parameters, namely the mean and variance associated with the normal distribution after a logarithm transformation. Their work is extended in May et al. (2000) and Spangler et al. (2004), where a third parameter (called *location parameter*) is added to a lognormal distribution. Both papers compare various strategies that estimate the location parameter. Using the same data set as the one in Strum et al. (2000), May et al. (2000) show that the skewness of data is an effective indicator that identifies the best estimation strategy. They also observe that when the skewness of data is small, the two-parameter lognormal models outperform the three-parameter lognormal models (i.e., the one with a location parameter). Spangler et al. (2004) suggest using a properly chosen order statistics for estimating the location parameter in a lognormal distribution. Both simulated and real surgical data (again, with single CPT codes) are used to test different estimation strategies in Spangler et al. (2004).

Surgeries consisting of exactly two CPT codes are the focus of Strum et al. (2003). Treating permutations of the same CPT codes as different combinations of CPT codes, Strum et al. (2003)

perform Shapiro-Wilk goodness-of-fit tests to examine the fitness of the lognormal and normal distributions for each combination of CPT codes. They conclude that lognormal distributions provide a better fit. Building on this result, Strum et al. (2003) apply logarithm transformations to normalize surgical case durations prior to conducting hypothesis testings with linear models. Their hypothesis tests show that permutations of CPT codes do not affect the accuracy of predictions of surgery case durations. Their results confirm that CPT code is the most important factor when predicting surgical case durations. Anesthesia types, emergency status, patient ages and surgery departments are also found to be relevant factors.

Although the importance of CPT codes in predicting surgical case durations has been noted for at least a decade, Strum et al. (2003) present the only work in the existing literature that uses CPT codes as explanatory variables for surgeries containing more than one CPT code. The limitation of their approach can be understood as follows. Their method provides a prediction only for surgeries consisting of exactly two CPT codes. When applying their method, a sufficient number of surgeries with the *same combination* of CPT codes must exist in the data samples. This requirement limits the application of their approach even for surgeries consisting of exactly two CPT codes. Moreover, it is difficult to extend Strum et al. (2003)'s distribution-fitting approach to surgeries with three or more CPT codes, due to lack of historical data. — Although there will be plenty of surgical cases with three or more CPT codes, there are not that many cases with exactly the same combination of three or more CPT codes.

In this paper, we propose two regression models to predict surgical case durations. Different from previous regression-based approaches, our models explicitly include CPT codes as the explanatory variables. The proposed models can be applied in general situations where a surgery can have any number of CPT codes and any combination of CPT codes. To the best of our knowledge, this paper is the first that develops a systematic approach to predict surgery case durations based on multiple CPT codes.

Utilizing CPT code information is not a trivial matter because (i) many CPT codes only appear in conjunction with others and thus do not have any predictive power on their own; (ii) combination of CPT codes varies from surgery to surgery. Incorporating CPT information in our regression models hinges upon constructing a suitable design matrix of existing CPT codes, which frees us from relying on the occurrence of the same combination of CPT codes in historical data. The main challenge of constructing such a suitable design matrix is that naively constructed design matrix is usually ill-conditioned (i.e. singular). We devise a construction procedure to obtain a nonsingular, well-conditioned design matrix for our regression models. Our procedure carefully sifts out those

CPT codes without any predictive power while retaining useful information as much as possible.

We compare our two regression-based models with three benchmark methods, one uses a log-normal distribution for prediction and the other two involve making predictions based on sample means. We measure the models' performances in terms of both mean squared errors (MSE) and mean relative absolute errors (MRAE). These performance measures show that our proposed models make more accurate predictions of surgical case durations than the benchmark methods although the magnitude of improvement varies for different service departments.

The rest of the paper unfolds as follows. In Section 2, we describe the surgical data set with which we establish our prediction models and validate our approaches. Details of our predictive models are presented in Section 3. We compare our predictions to that of three benchmark approaches in Section 4. Section 5 concludes the paper.

## 2 Data Set

Our surgical data set is from a large teaching hospital in central Texas. The data set consists of 48,714 surgical cases from 10/1/2004 to 3/31/2008. It involves 25 operating rooms (OR), 115 surgeons and 19 service departments. Variables collected include surgery date, operating room number, surgeon's name initial, the date and time at which a patient was admitted into an OR (*pt\_in*), surgery preparation began (*prep\_pos*), surgery began (*incision*), surgery ended (*closure*), dressing ended (*dress\_end*) and the patient left OR (*pt\_out*), as well as the CPT code(s), which accurately describe the surgical procedures performed. Here is an example of a surgery. The surgery was performed on a weekday in March 2008. According to the records, the patient entered the operating room at 11:34am. Preparation for surgery started at 11:58am. The surgeon made the first incision at 12:03pm and closed the patient up at 13:10pm. By 13:20pm the patient was completed dressed. He/she was transported out of the operating room at 13:28pm. Three CPT codes were performed during the period from 12:03pm (*incision*) to 13:10pm (*closure*). They are 25607, 64415, and 76942. In our data set, a surgical case could include as many as eight CPT codes.

Among the various segments of time included in a surgery, we are most interested in the *surgical time* (the duration from *incision* to *closure*). This is the time during which surgical actions take place. The CPT codes prescribed for a surgery are carried out during this time, and hence have a direct impact on the length of this time. The surgical case durations we study in this paper refer to these surgical times.

**Remark.** [We recognize that other durations, for example, the *total time* (the duration from *pt\_in* to *pt\_out*), can also be of interest to practitioners and researchers. In addition to surgical times, Strum et al. (2003) also examine total times in their work. Although our focus is on surgical times, our proposed methodology can be easily adapted to the modeling of total times. The adaptation, however, adds little insight. In order to avoid repetitions, we present our methods in the context of surgical times.]

Before we establish our predictive models using the historical data, a data cleaning action is performed to eliminate data records that are deemed “invalid.” The following considerations are used to identify invalid records: (i) A record should have a starting time entry and an ending time entry to calculate a time duration. When a data record lacks any one of the two entries, it is incomplete and thus not used. (ii) The starting time should be earlier than the ending time. (iii) The duration should not be unreasonably long; for instance, 36 hours would be considered anomalous for our contact hospital given the nature of their operations. If these conditions are not met, a record is invalid and removed from the data set. Tables 1 and 2 provide summary information regarding our surgical data after cleaning actions. From Table 1, one observes that data cleaning only eliminates a tiny portion of the data records (about 0.7% of the original data with a total of 48,714 cases). One also observes that a large portion of the surgical cases include no more than two CPT codes. However, Strum et al. (2003)’s approach can not be readily applied here because those cases with two CPT codes do not necessarily share two common CPT codes. In our data set, there were 11,771 combinations of CPT codes among 48,373 valid cases. That is, there were, on average, about 4.1 cases with the same combination of CPT codes. Furthermore, even though the cases with more than two CPT codes are in minority among the total of forty-eight thousand plus cases, the absolute number of those cases (totaling 8,754) is remarkable. The importance of making accurate duration predictions of these cases with at least three CPT codes should not be understated.

Table 2 provides departmental statistics of the surgical cases. The surgical cases in our data set were performed by 19 service departments, each in charge of a specialty area, for example, orthopedics, oncology, etc. Each department is represented by an acronym (consisting of two or three letters) commonly used and readily recognizable in medical profession; hence we skip the explanation of these acronyms. Although CPT codes associated with surgeries performed by different service departments often differ, we find that some CPT codes are shared across various departments. This is not surprising since two surgical cases serving different purposes could have a common set of surgical actions, which is represented by a common CPT code.

Number of CPT codes included in a surgery ( $k$ )	Number of valid cases
$k = 1$	29,039
$k = 2$	10,580
$k = 3$	4,065
$k = 4$	2,172
$k = 5$	1,182
$k = 6$	574
$k = 7$	366
$k = 8$	395
SUM	48,373

Table 1: Number of valid cases with exactly  $k$  ( $k = 1, \dots, 8$ ) CPT codes after data cleaning.

For each department, Table 2 lists the number of valid cases, the number of CPT codes, and the number of CPT combinations performed by that department. Note that a CPT combination is a set of CPT codes that appear together in a surgical case. Permutations of the same set of CPT codes are treated as the same CPT combination because permutations do not have any significant impact on surgical case durations (see Strum et al. 2003). The goal of this research is to predict surgical case durations based on the CPT codes included in a surgery. Table 2 roughly outlines the size of the problem we are dealing with. It is also clear from Table 2 that the surgical case load distributions across various departments are uneven; most departments have performed over one thousand surgeries over the three-and-half-year period, whereas a few departments have performed fewer than one hundred cases.

### 3 Solution Approaches

Surgical case durations are predicted for each service department separately. The reason is threefold. Firstly, each service department handles their own surgery schedules. Secondly, the CPT codes that describe the surgical procedures, despite certain degree of sharing, are by and large different across the service departments. Thirdly, service departments are found to be a relevant factor that affects the prediction of surgical case durations (see Strum et al. 2003).

The need of department-specific predictions further renders the existing lognormal distribution based approach less effective. If we are to estimate the lognormal distribution for a given CPT combination, we hardly have enough data points in the sample. To see this, one can simply compare the number of valid cases and the number of CPT combinations performed by each department in Table 2. The average number of cases per CPT combination ranges from 1 to 7.54 among the 19 service departments; apparently 7 cases per CPT combination are not enough data points for

Dept.	# of valid cases performed by a dept.	# of CPT codes performed by a dept.	# of CPT combinations performed by a dept.
NS	1,060	207	312
ORT	8,606	928	2,308
TPT	1,500	100	199
URO	5,223	489	1,108
CT	1,825	194	606
THO	721	241	502
UMC	1,390	280	338
GEN	8,386	656	1,507
ONC	4,755	579	1,493
GYN	4,726	366	906
ORA	555	181	206
PLA	3,118	736	1,486
EYE	89	54	54
VAS	1,549	309	818
PDS	2,489	397	636
ENT	1,960	363	544
OTH	65	23	20
POD	354	85	113
RAD	2	3	2

Table 2: Number of valid cases, number of CPT codes, and number of CPT combinations performed by each service department after data cleaning.

distribution estimation.

Figure 1 shows the histograms of data for three departments (CT, UMC and ORA), where the data deviate significantly from lognormal distributions. In particular, the case durations in service department CT has a bi-modal distribution, which cannot be well approximated by either the lognormal distribution or the normal distribution.

What we propose here is a regression-based approach, which made no assumption on normality or normality after logarithm transformation. Our models explicitly use CPT codes describing surgical procedures as explanatory variables. This allows the specific knowledge regarding a surgical procedure to be incorporated. We would like to note that in the current research we consider only CPT codes, while ignoring other possible covariates, since CPT codes are recognized as the most important factor in representing surgical case durations in existing literature. Our later numerical results indeed demonstrate enough benefit of our research undertaking. We do acknowledge that considering both CPT codes and other important covariates (such as surgeons or anesthesia types) could potentially further improve the prediction of surgical case durations. But doing so will require a different data set, and is thus out of the scope of this paper.

In the sequel, we will first present two regression models that predict the surgical time (the duration from *incision* to *closure*). We then describe a singularity problem encountered in applying

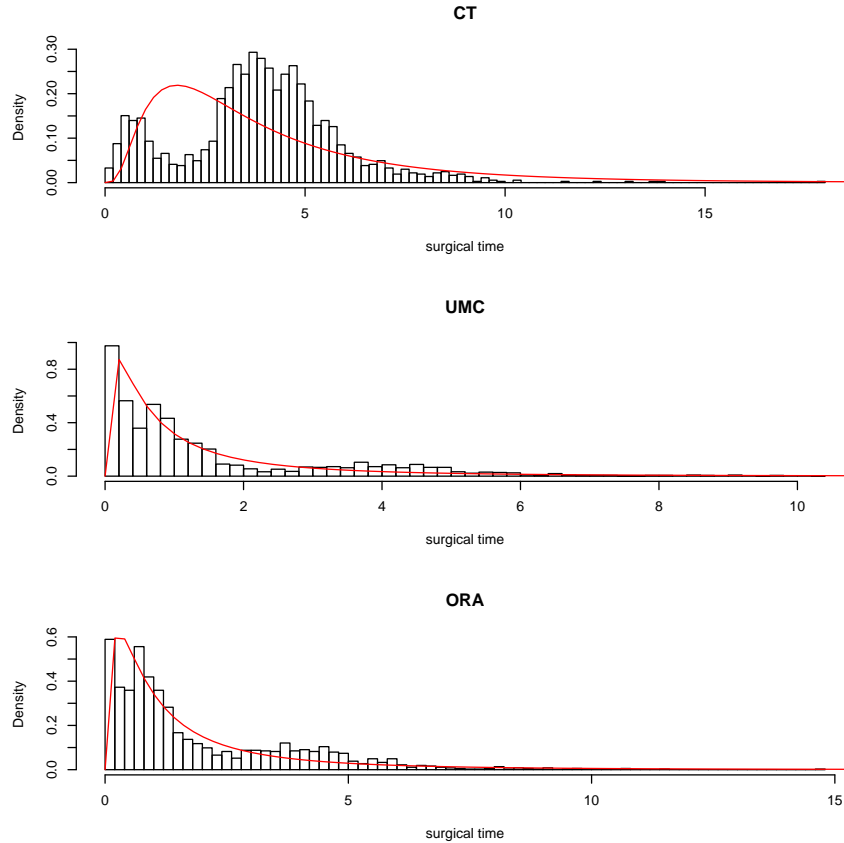


Figure 1: Histograms and best-fit lognormal densities of the surgical time for three service departments. The unit for the horizontal axis is in hours.

these regression models. In the rest of Section 3 we propose systematic procedures that address the problem.

### 3.1 Regression models

Since CPT codes describe specific surgical actions undertaken in a surgery, the surgical time can naturally be considered as the summation of all the component surgical actions. Suppose there are  $n$  surgical cases performed by a given service department involving a total of  $m$  CPT codes. Denote by  $y_i$  the surgical time of case  $i$ . We introduce here an indicator variable,  $x_{ij}$ , of the inclusion of the  $j$ th CPT code in the  $i$ th surgical procedure. In other words, when CPT code  $j$  shows up in surgical case  $i$ ,  $x_{ij} = 1$ ; otherwise  $x_{ij} = 0$ . Denote by  $\beta_j$  the expected time of performing the surgical action specified by CPT code  $j$ . We have the following model to describe the surgical time:

$$y_i = \sum_{j=1}^m x_{ij}\beta_j + e_i, \quad i = 1, \dots, n$$

where  $y_i$  is the summation of the expected times associated with all the CPT codes involved in case  $i$ , plus  $e_i$ , which is the residual error of case  $i$  that cannot be modeled by the expected times. Residual error  $e_i$  is assumed to be a zero-mean random variable.

The above model can be expressed in a matrix form as:

$$Y = X\beta + e \tag{1}$$

where  $Y = (y_1, \dots, y_n)'$  is the  $n \times 1$  vector of surgical times,  $\beta = (\beta_1, \dots, \beta_m)'$  is the  $m \times 1$  vector of the expected times associated with  $m$  CPT codes,  $X = (x_{ij})$  is the  $n \times m$  design matrix, representing the inclusion of CPT codes in surgical cases.

Equation (1) represents a typical linear regression model. Once surgeries are performed,  $Y$  and  $X$  are known, and  $\beta$  is the one to be estimated from historical data. Denote by  $\hat{\beta}$  the estimate of  $\beta$ , and  $\hat{\beta}$  will be used in future predictions. Since we do not restrict the sign of our estimates, it is possible that we obtain negative estimates of the expected times for certain CPT codes. The predicted surgical time could still be positive because it is determined by the combination of the comprising CPT codes. The negativity can be completely avoided by adding a non-negativity constraint on the  $\beta_j$ 's. We did not impose this constraint in our implementation of the regression models because negative values rarely appear in our analysis. More details and explanations are reported at the end of Section 3.

In order to predict the surgical time of a new case  $z$ , one needs to look at the CPT codes to be performed in this case and create a design vector  $x_z$  by assigning “1” or “0” to the corresponding  $x_{zj}$  for  $j = 1 \dots m$ . Then, calculating the inner product of this design vector  $x_z = (x_{z1}, \dots, x_{zj}, \dots, x_{zm})$  with the estimates  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_m)'$  gives the predicted surgical time of the new case. Precisely, let  $Y_z^{new}$  denote the surgical time of the new case, then the linear regression model predicts that  $\widehat{Y}_z^{new} = x_z \hat{\beta}$ .

The above model is flexible and easy to use in predicting surgical times composed of any number of CPT codes. Suppose one extra set of surgical actions is added to a series of existing actions, then the surgical time will be simply increased by the length of the corresponding actions described by the additional CPT code(s). The model sets no restrictions on how many CPT codes a surgical procedure can include or what CPT combinations should appear.

Next we present an alternative model, motivated by the arguments in existing literature that surgery data are better fit by a lognormal distribution (Strum et al. 2000, Strum et al. 2003, May et al. 2000, and Spangler et al. 2004, among others). These arguments help legitimize the use of logarithm transformation to normalize surgical procedure times. In light of this, we take logarithm

transformations of surgical times before fitting them to a linear regression model. Consequently, our second model reads as

$$\log(Y) = X\beta + e \quad (2)$$

where  $\log(Y) \equiv (\log(y_1), \dots, \log(y_n))'$ , an  $n \times 1$  vector, and  $e$  is a vector of zero-mean residuals. We refer to equation (2) as a log-regression model. Again,  $Y$  and  $X$  are known from historical data, and  $\beta$  is to be estimated. The estimate  $\hat{\beta}$  will be used for future predictions, in a similar fashion as in the linear regression model explained above. For a new case  $z$  with corresponding design vector  $x_z$ , the prediction of the surgical time  $Y_z^{new}$  is given by  $\widehat{Y}_z^{new} = \exp(x_z \hat{\beta})$ .

Compared to the linear regression model (1), the log-regression model (2) is less intuitive in terms of its practical interpretation. Note that a surgical case comprises of a series of component procedures (each is represented by a CPT code). The linear regression model implies that the surgical time is the summation of the times associated with the component procedures, while the log-regression model suggests that the surgical time is the product of the exponentials of the times associated with the component procedures. The advantage of the log-regression model is that its prediction is always positive. Our numerical results (presented in Section 4) show that both models perform well.

A point worth noting is that we do not make distribution assumptions in (1) and (2). By using the method of least squares to fit model (1), we find the best linear predictor of the surgical time. Although model (2) is motivated by log-normal distribution arguments, least squares fitting of the model can be considered as finding the best linear prediction of the log surgical time. The logarithmic transformation is simply used as a device to ensure a positive prediction.

**Remark.** If the duration of interest is *total time*, namely the duration a patient spends in an OR, the models in (1) and (2) only need to be slightly modified. Noticing that the total time is the addition of the surgical time and the pre- and post-surgery processing times, we can add an intercept term  $\beta_0$  to both models in (1) and (2). As such, the models for the total time read as:

$$Y = \beta_0 \cdot \mathbf{1}_n + X\beta + e \quad (3)$$

and

$$\log(Y) = \beta_0 \cdot \mathbf{1}_n + X\beta + e \quad (4)$$

where  $Y$  now represents the total time,  $\log(Y)$  follows the same notation as in model (2), and  $\mathbf{1}_n$  is an  $n \times 1$  vector whose elements are all 1's. In model (3),  $\beta_0$  represents the expected time consumed collectively by all the pre- and post-surgery actions. Similar meaning applies to  $\beta_0$  in

model (4) which is after a logarithm transformation. The inclusion of extra durations bring in additional variability, which is absorbed into the residual error  $e$  in the above models. After models (3) and (4) are fit using the training data, predictions of the total time can be easily computed. Let  $x_z$  denote the design vector for a new case, the corresponding total time  $Y_z^{new}$  can be predicted using  $\widehat{Y}_z^{new} = \hat{\beta}_0 + x_z \hat{\beta}$  for model (3) and  $\widehat{Y}_z^{new} = \exp(\hat{\beta}_0 + x_z \hat{\beta})$  for model (4). In the interest of conciseness, we report our proposed procedures in the context of surgical times.

### 3.2 Singularity of design matrix $X$

After establishing the regression models (1) and (2), if the design matrix  $X$  is of full rank, we can estimate  $\beta$  in the regression models through a standard least-squares estimation (Weisberg, 2005). Specifically, for model (1)

$$\hat{\beta} = (X'X)^{-1}X'Y; \tag{5}$$

and for model (2), one needs to simply replace  $y_i$  with  $\log(y_i)$  and  $Y$  with  $\log(Y)$ . The reason that a fully ranked  $X$  is required is because of the inversion on  $X'X$ .

Whether the design matrix  $X$  is of full rank, however, depends on how it is constructed. If we list all the CPT codes performed by a service department and naively use this list to construct  $X$ , we will obtain an ill-conditioned  $X$ . As a result,  $X'X$  is not invertible and  $\beta_i$ 's cannot be estimated. Consequently, the expected times associated with the corresponding CPT codes cannot be estimated.

Consider for example three CPT codes,  $A$ ,  $B$ , and  $C$ . (Note that we use capital letters to denote CPT codes for the sake of simplicity although an actual CPT code is a five-digit number.) Assume that the expected times for the three CPT codes are  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ , respectively. Suppose that there are only three surgical cases: the first case uses all three CPT codes, the second case uses CPT codes  $B$  and  $C$ , and the third case uses only CPT code  $A$ . Then, a naively constructed design matrix  $X$  using the CPT code list is

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

which is singular.

The singularity in the above illustration is caused by the co-appearance of CPT codes  $B$  and  $C$ . From the surgical cases performed and times measured, we can obtain only the combined time of these two CPT codes, not the time associated with the individual CPT code  $B$  or  $C$ . In general,

to check whether a certain CPT code  $B$  always appears in conjunction with another code  $C$ , we can perform a simple test as follows: count the number of appearances of CPT codes  $B$  and  $C$  in the surgical cases within a given service department; suppose both appear, for instance,  $h$  times. Then, count the number of CPT codes  $B$  and  $C$  appearing together (we call this CPT combination  $BC$ ). If  $BC$  also appears  $h$  times, then it implies that CPT codes  $B$  and  $C$  always appear together in conjunction with each other. This appearance pattern will result in a singular design matrix. Furthermore, a CPT code can appear in conjunction with different CPT codes in different cases. Therefore, one needs to exercise extra care when constructing a design matrix. Recall that we have to deal with a large number of CPT codes for each department. Next, we propose a systematic procedure that thoroughly and efficiently sifts out CPT codes/combinations that cause singularity, without losing information.

### 3.3 Grouping CPT combinations

In order to avoid singularity, CPT codes that always appear together should be treated as a whole as if they formed a new CPT code. For instance, in the singularity example above, instead of attempting to estimate individually the times associated with  $A$ ,  $B$ , and  $C$ , one should only try to estimate the times associated with a single CPT code  $A$  and a CPT combination  $BC$ .

In light of this, we need to group CPT codes and combinations in our data set appropriately. The purpose of grouping is to establish the set of single CPT codes whose execution times can be estimated, the set of two-code CPT combinations whose combined time can be estimated, the set of three-code CPT combinations whose combined time can be estimated, and so on. A full-rank design matrix can then be constructed based on the grouping results. We will explain the construction of a design matrix in the next subsection.

Before we present our detailed grouping procedure, we introduce the concept of *code length*, which is defined as the number of component CPT codes in a CPT combination. Denote by  $k$  the code length of a CPT combination. In our data set, the largest  $k$  is eight. In the sequel we only illustrate implementation details for  $k$  up to eight, although the general procedure applies to any value of  $k$ . Given the fact that our data set covers nearly 50 thousand cases over three and half years, the scenario in which  $k$  could be greater than eight should rarely happen in reality. The grouping procedure is as follows.

- First, construct  $k = 8$  empty sets  $S_1, \dots, S_k, \dots, S_8$ , where  $S_k$  will hold the grouping results for CPT combinations of length  $k$ .

- Repeat the following for  $k = 1, 2, \dots, 8$ 
  - Identify all the surgical cases with exactly  $k$  CPT codes. Put them in  $S_k$ . If there are no such cases, we have finished selecting the CPT combinations of length  $k$ , so go to the next value of  $k$ .
  - For each CPT combination of length  $k$  in  $S_k$ , determine whether it is “distinctive.” We now describe how the distinctiveness of a CPT combination is determined. A CPT combination of length  $k$  can be decomposed into a number of CPT codes or code combinations of length 1 to length  $k - 1$ . For instance, a CPT combination  $ABC$  of length 3 can be decomposed into three single CPT codes of length 1,  $A$ ,  $B$ ,  $C$ , or a CPT combination of length 2 plus a single code; there are three possibilities, i.e.,  $AB$  and  $C$ , or  $AC$  and  $B$ , or  $BC$  and  $A$ . If there exists a decomposition scheme in which all the decomposed component codes or code combinations can be found in sets  $S_1$  to  $S_{k-1}$ , then the CPT combination of length  $k$  is not distinctive; otherwise it is.
  - Remove all the non-distinctive CPT combinations of length  $k$  from  $S_k$ .

In the above procedure, the step of determining the distinctiveness of a CPT combination is relatively involved. For  $k = 1$ , it is straightforward since there is no set  $S_0$ , all single CPT codes automatically satisfy the distinctiveness condition. For  $k = 2, \dots, 8$ , we have to go through all possible decomposition schemes of a CPT combination of length  $k$ . The larger the  $k$ , the more complicated a decomposition process becomes.

Table 3 helps sort out the decomposition schemes for  $k = 2, \dots, 8$ . To understand the notation in the table, take  $k = 4$  as an example. The entry of  $4 = 3 + 1$  means that a CPT combination of length 4 can be decomposed into a CPT combination of length 3 and one of length 1 (a single code); the next lines of  $4 = 2 + 2$ ,  $4 = 2 + 1 + 1$ , and  $4 = 1 + 1 + 1 + 1$  mean that the same CPT combination can also be decomposed into two CPT combinations of length 2, or a CPT combination of length 2 plus two single codes, or four single codes, respectively. Collectively, those are all the possible decomposition schemes for a CPT combination of length 4.

Recall that Strum et al. (2003) found that permutations of component CPT codes did not significantly affect surgical case durations. For this reason, we do not consider permutations of a CPT combination any different than the original CPT combination. Our definition of the distinctiveness of a CPT combination is based on the decomposition of the CPT combination, not permutations. Another note is that the above grouping procedure can be applied to any surgical

$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
2 = 1+1	3 = 2+1	4=3+1	5=4+1	6=5+1	7=6+1	8=7+1
	3 = 1+1+1	4=2+2	5=3+2	6=4+2	7=5+2	8=6+2
		4=2+1+1	5=3+1+1	6=4+1+1	7=5+1+1	8=6+1+1
		4=1+1+1+1	5=2+2+1	6=3+3	7=4+3	8=5+3
			5=2+1+1+1	6=3+2+1	7=4+2+1	8=5+2+1
			5=1+1+1+1+1	6=3+1+1+1	7=4+1+1+1	8=5+1+1+1
				6=2+2+2	7=3+3+1	8=4+4
				6=2+2+1+1	7=3+2+2	8=4+3+1
				6=2+1+1+1+1	7=3+2+1+1	8=4+2+2
				6=1+1+1+1+1+1	7=3+1+1+1+1	8=4+2+1+1
					7= 2+2+2+1	8=4+1+1+1+1
					7= 2+2+1+1+1	8=3+3+2
					7= 2+1+1+1+1+1	8=3+3+1+1
					7= 1+1+1+1+1+1+1	8=3+2+2+1
						8=3+2+1+1+1
						8=3+1+1+1+1+1
						8=2+2+2+2
						8=2+2+2+1+1
						8=2+2+1+1+1+1
						8=2+1+1+1+1+1+1
						8=1+1+1+1+1+1+1+1

Table 3: Decomposition schemes of a CPT combination of length  $k$ .

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	Total
Number of cases with $k$ codes in ENT	1297	482	111	37	19	7	3	4	1960
Number of CPT combinations of length $k$	182	205	92	32	19	7	3	4	544
Number of distinctive CPT combinations of length $k$	182	119	45	14	12	7	3	4	386

Table 4: Summary of CPT combinations in Department ENT

data set but in this research we apply them to the data of individual service departments due to the department-specific approach we undertake in predicting the surgical case durations.

As an illustration, we present Table 4, which summarizes the the number of valid cases, the number of CPT combinations, and the number of distinctive CPT combinations for the service department of ENT. There are a total of 1,960 valid cases and 544 CPT combinations (including single CPT codes). Among the 544 CPT combinations, 386 of them are distinctive according to the aforementioned definition, and others can be decomposed into components found in the sets of shorter code length. For instance, there are 205 CPT combinations of length 2, but 86 of them can be decomposed into two single CPT codes that are both present in  $S_1$ . That leaves 119 ( $= 205 - 86$ ) distinctive CPT combinations of length 2 in  $S_2$ . Therefore, the size of  $S_2$  becomes 119 after our grouping procedure is applied.

### 3.4 Constructing a design matrix

Constructing a design matrix is to assign “1” or “0” to each element  $x_{ij}$  in matrix  $X$ . Recall that when previously introduced, the first index  $i$  is the case index, ranging from 1 to  $n$ , and the

second index  $j$  is the CPT code index, ranging from 1 to  $m$ . After applying the grouping procedure described in Section 3.3, we will estimate the expected times  $\beta_j$  not only for single CPT codes but also for distinctive CPT combinations of length  $k \geq 2$ . So the value of  $m$  depends on the number of distinctive CPT combinations (including single CPT codes) in a given data set. For example, let's take the data in Table 4 for illustration. If surgeries with code length up to 8 are to be included in the regression model, then  $m = 386$ . But if only surgeries with the single CPT codes and those with code length of 2 are to be included, then  $m = 301 (= 182 + 119)$ . Suppose that we include all the surgeries with code length up to 8. We should, then, aggregate all the distinctive CPT combinations (sets  $S_1$  to  $S_8$ ) into a set  $S \equiv \bigcup_{k=1}^8 S_k$ . If there are a total of  $m$  elements in  $S$ , a row vector of matrix  $X$ ,  $x = (x_{i1}, \dots, x_{ij}, \dots, x_{im})$ , has a one-to-one correspondence to the  $m$  elements in  $S$ .

For convenience, we order the surgical cases based on the number of CPT codes they have, namely that first comes the surgical cases with a single CPT code, followed by the surgical cases with exactly two CPT codes, and then followed by the cases with exactly three CPT codes, and so on. So eventually, surgical cases 1 to  $i_1$  has a single CPT code, cases  $i_1 + 1$  to  $i_2$  has two CPT codes,  $\dots$ , and cases  $i_7 + 1$  to  $i_8$  has eight CPT codes, where  $1 \leq i_1 \leq \dots \leq i_8 = n$ .

The basic idea of constructing a design matrix is that for each  $i = 1, \dots, i_8$ , take the corresponding surgical case and match the CPT codes it has with the distinctive CPT combinations in  $S$ . If a match is found, then the corresponding  $x_{ij}$  will be set to "1"; otherwise  $x_{ij}$  will be set to "0". We here assume that the set  $S$  is well maintained and timely updated using our grouping procedure. So the finding of a match is guaranteed.

Despite the simplicity of this idea, certain complexities have to be dealt with. For surgical cases with a single CPT code (cases 1 to  $i_1$ ), the construction procedure is just like what the basic idea describes, except that one need not search the set of  $S$  but only  $S_1$ . For surgical cases with two CPT codes (cases  $i_1 + 1$  to  $i_2$ ), the two codes could appear as a CPT combination of length 2 or they may have appeared as two single CPT codes. What one needs to do is to search first the set of  $S_2$  in order to check if there is a match for a CPT combination of length 2, and if not, then search  $S_1$  for the matches of the two single CPT codes. Depending on the outcome of the search, the appropriate  $x_{ij}$  can be set to 1. For surgical cases with  $k \geq 3$  CPT codes (cases from  $i_2 + 1$  onward), one needs to search for matches in different sets from  $S_k$  to  $S_1$ , similarly as one does for surgical cases with two CPT codes. Because there are many different ways of decomposing a CPT combination when  $k$  gets large, Table 3 is a good reference that can guide the search process.

In addition to searching for matches from all possible schemes of decomposition, one more

Case #	CPT codes performed in a case
1	A
2	C
3	AB
4	BC
5	ABC
6	ABD
7	ABCD
8	ABCDEF

Table 5: CPT codes relating to the cases in the design matrix example.

complexity arises for surgical cases having three or more CPT codes. To understand this, take a surgery case with three CPT codes as an example. Suppose that the CPT codes prescribed for the surgery are  $A$ ,  $B$  and  $C$ . Also suppose when searching the set of  $S_3$ , we do not find any matches; and when searching the set of  $S_1$ , not all three of the single codes found their matches, either. Then, we need to search  $S_2$  for possible matches of a CPT combination of length 2. Doing so could give us multiple matches: for example, we could have  $AB$  in  $S_2$  while  $C$  in  $S_1$ , this is one match; or  $AC$  in  $S_2$  while  $B$  in  $S_1$ , this is another match. If both matches are found, the surgical case in question can be used to estimate both the expected times of  $AB$ ,  $C$ , and the expected times of  $AC$ ,  $B$ , unless one has profound prior knowledge suggesting otherwise. As a matter of fact, in order to extract the most information from this surgical case, its duration should be taken into account when we estimate both the expected times of  $AB$ ,  $C$ , and the expected times of  $AC$ ,  $B$ . To do so properly, we should include this surgical case twice in our design matrix. One inclusion represents the decomposition  $AB+C$ , and the other  $AC+B$ . In order to account for duplicate use of the same data, caused by multiple inclusions of a single surgical case, we use the weighted least-squares approach by applying weights that are inversely proportional to the number of inclusions.

To illustrate, consider the following example. Suppose a data set contains only eight surgical cases. The CPT codes prescribed for each case are listed in Table 5.

Apparently there are six CPT codes ( $A$ ,  $B$ ,  $C$ ,  $D$ ,  $E$  and  $F$ ) ever performed. It is straightforward to verify that a naively constructed design matrix that assigns “0” and “1” to each case based on their inclusion of each of the six CPT codes is singular. Therefore, the design matrix should be constructed differently. Applying the grouping procedure from Section 3.3, we obtain the following sets  $S_1 = \{A, C\}$ ,  $S_2 = \{AB, BC\}$ ,  $S_3 = \{ABD\}$ , and  $S_6 = \{ABCDEF\}$ . Aggregating  $S_1$ ,  $S_2$ ,  $S_3$  and  $S_6$  generates  $S = \{A, C, AB, BC, ABD, ABCDEF\}$ , implying that  $m = 6$ . That is, there are six columns in the design matrix. Next we construct each row of the design matrix by including each of the surgical cases in the data set.

$$X = \begin{array}{cccccc|l}
& A & C & AB & BC & ABD & ABCDEF & \\
\left[ \begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array} \right. & \begin{array}{l}
\text{Case 1} \\
\text{Case 2} \\
\text{Case 3} \\
\text{Case 4} \\
\text{Case 5 (first inclusion)} \\
\text{Case 5 (second inclusion)} \\
\text{Case 6} \\
\text{Case 7} \\
\text{Case 8}
\end{array}
\end{array}$$

Figure 2: Design matrix example. The code or code combination on the top indicate the columns corresponding to  $A, C, AB, BC, ABD, ABCDEF$ , respectively, and the texts on the right side of the matrix identify the corresponding cases.

For Case 1, apparently,  $x_{11} = 1$  and all other entries in the first row of the design matrix  $X$  are zeros since the CPT code used by Case 1 matches only the first element in  $S$ . Following the same reasoning, for Case 2,  $x_{22} = 1$ ; for Case 3,  $x_{33} = 1$ ; and for Case 4,  $x_{44} = 1$ . For Case 5,  $ABC$  can have two different decompositions,  $C + AB$  or  $A + BC$ . Both decompositions are possible in  $S$ . Therefore we include Case 5 twice (occupying two rows) in the design matrix: for the fifth row,  $x_{51} = 1$  and  $x_{54} = 1$  (corresponding to  $A + BC$ ); and for the sixth row,  $x_{62} = 1$  and  $x_{63} = 1$  (corresponding to  $C + AB$ ). Because Case 5 is included twice, Cases 6, 7 and 8 will then correspond to rows 7, 8 and 9 (instead of rows 6, 7 and 8) of the design matrix  $X$ , respectively. For Case 6,  $x_{75} = 1$ ; for Case 7, since  $ABCD$  can be decomposed into  $C + ABD$ ,  $x_{82} = 1$  and  $x_{85} = 1$ ; for Case 8,  $x_{96} = 1$ . Ultimately, the design matrix  $X$ , which is of full rank, looks like the matrix presented in Figure 2. Note that there are 8 surgical cases in the example while the resulting design matrix has 9 rows.

Below we outline a general procedure for the construction of a design matrix assuming that the set  $S$  has already been obtained. Let  $m$  be the total number of elements in the set  $S$ , and  $n$  the total number of surgical cases in the data set.

- (1) Set the number of columns in the design matrix to  $m$ . Order the elements in the set  $S$  from 1 to  $m$ , and use them to label the columns of the design matrix.
- (2) Order all the surgical cases from 1 to  $n$ .
- (3) Set  $i = 1$  and  $r = 1$ .

- (4) Decompose the  $i_{th}$  surgical case using the elements in the set  $S$ . Let  $d_i$  be the number of possible decompositions. Order the possible decompositions from 1 to  $d_i$ . Set  $c = 1$ .
- (5) For  $j = 1, \dots, m$ , use  $x_{rj}$  to denote the value of the entry at the intersection of the  $r_{th}$  row and the  $j_{th}$  column. Set  $x_{rj} = 1$  if the  $c_{th}$  decomposition of the  $i_{th}$  case uses the  $j_{th}$  element of the set  $S$ . Otherwise,  $x_{rj} = 0$ .
- (6) If  $c = d_i$  then set  $r = \sum_{t=1}^i d_t + 1$  and go to Step (7). Otherwise set  $c = c + 1$ ,  $r = r + 1$  and go to Step (5).
- (7) If  $i = n$  then the design matrix is completed. Otherwise set  $i = i + 1$  and go to Step (4).

Next we illustrate the application of our design matrix using data from the service department ENT. We use  $S = (S_1, S_2, S_3, S_4)$ , our  $m = 360$  ( $=182+119+45+14$ ), and  $n = 1,927$  ( $=1,297+482+111+37$ ). The corresponding design matrix  $X$  is of dimension  $1,950 \times 360$  rather than  $1,927 \times 360$  because of repeated inclusions of certain cases for the reasons explained earlier. We then proceed to estimate the  $\beta$ 's associated with the 360 distinctive CPT combinations (using equation (5)); they are the expected times for performing the corresponding CPT combinations. Figure 3 presents the histogram for the 360 values of  $\hat{\beta}$ 's. Most of these CPT combinations have an estimated time in the range of (0, 10) hours. The vertical line is the mean of the times of the 360 distinctive CPT combinations, which is about 1.532 hours. Almost half of the distinctive CPT combinations used in department ENT takes fewer than one hour to complete. We also observe from Figure 3 that a very small portion of distinctive CPT combinations have a negative time estimate. In fact, 4 out of these 360 (1.11%) CPT combinations have a negative time estimate. To explain why we may have negative estimates, consider the following example of two surgical cases: Case 1 includes CPT codes  $A$  and  $B$ , while Case 2 only uses CPT code  $A$ . When surgeons actually perform these surgical cases, the surgical time of Case 1 could be shorter than that of Case 2. When this happens, the estimated time of CPT code  $B$  becomes negative. This negativity rarely happens, as evident from the ENT departmental data (data from other departments generate the same conclusion). Moreover, the code  $B$  is likely to appear together with another CPT code and thus still gives a positive prediction of the surgical time. We are confident that the rare appearance of negativity does not cause our prediction of surgical case durations to go off the marks.

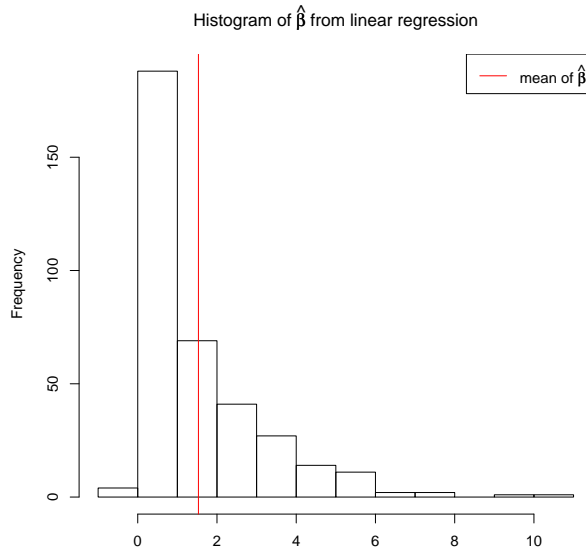


Figure 3: Histogram of  $\hat{\beta}$ 's from the linear regression model for department ENT. Unit for the horizontal axis is in hours.

## 4 Prediction and Comparison

### 4.1 Construction of training and test data sets

According to the statistical literature (Witten and Frank, 2005; Mitchell, 1997), the typical protocol for validating empirical models is to split the original data set into a training set and a test set. Suppose the  $n$  surgical cases in the original dataset are divided as the  $n_t$  cases in the training set and the  $n_s$  cases in the test set, where  $n = n_t + n_s$ . The training data set is used to obtain  $\hat{\beta}$  based on equation (5); this is known as *model fitting*. After the model is fit, i.e., all  $\beta$ 's are estimated, one can use the model to make predictions of surgical case durations on the data records in the test set, which are never used in the model fitting process. Then, the predictions are compared with real measurements of surgical case durations in the test set. The differences between the predictions and the actual surgical case durations are good indications of how well a model works.

Suppose we would like to predict the duration for surgical case  $i$  in the test set. A design vector  $x_i$  for the case can be generated based on the set  $S$ , which is obtained after our grouping procedure is applied. Use the random variable  $Y_i^{new}$  to denote the surgical time of the case  $i$ . The predicted value is denoted as  $\widehat{Y}_i^{new}$ . The difference between the predictions and the actual surgical case durations is measured by two metrics: the mean squared errors (MSE) and the mean relative

absolute errors (MRAE). They are defined as:

$$MSE = \frac{1}{n_s} \sum_{i=1}^{n_s} \{y_i - \widehat{Y}_i^{new}\}^2 \quad \text{and} \quad MRAE = \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{|\widehat{Y}_i^{new} - y_i|}{y_i}$$

where  $y_i$  is the recorded duration of the  $i_{th}$  surgical case in the test set and  $\widehat{Y}_i^{new}$  is the predicted duration for the same case. MRAE characterizes how well a model makes prediction.

In this research, we assign two thirds of the historical cases to the training set and one third to the test set, i.e.,  $n_t \approx \frac{2}{3}n$  and  $n_s \approx \frac{1}{3}n$ , where “ $\approx$ ” is used because  $n_t$  and  $n_s$  need to be rounded off to the closest integer number. To avoid any systematic bias, the assignment of a case to one of the sets is randomly decided. Moreover, we repeat the assignment process 1,000 times, meaning that we randomly split the original data set 1,000 times, and consequently, we obtain 1,000 pairs of training/test sets. The performance measures MSE/MRAE, are then calculated 1,000 times using the 1,000 pairs of training/test sets. The MSE/MRAE values reported later in this section are the average of the 1,000 individual MSE/MRAE values.

## 4.2 Three benchmark methods

We compare our predictive models with three benchmark models below.

- Lognormal Model

This model assumes that the surgical time or the total time  $Y$  follows the lognormal distribution. It means that  $Y \sim \text{lognormal}(\mu, \sigma)$ , or equivalently,  $\log(Y) \sim \text{normal}(\mu, \sigma)$ , where  $\mu$  and  $\sigma$  are the parameters to be estimated. One can estimate them by using the data in the training set, such as:

$$\hat{\mu} = \frac{1}{n_t} \sum_{i=1}^{n_t} \log(y_i); \quad \hat{\sigma}^2 = \frac{1}{n_t} \sum_{i=1}^{n_t} (\log(y_i) - \hat{\mu})^2$$

Then, the surgical time for surgical case  $z$  in the test set is predicted using

$$\widehat{Y}_z^{new} = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right)$$

because that is the expectation of a lognormal model with parameters  $\mu$  and  $\sigma$  (Casella and Berger, 2001).

When using the lognormal benchmark model, we compute an estimated surgery length for all the surgeries in a department based on all the historical data in the training set for the same department. Ideally, we would like to find benchmark predictions for each surgery of a specific CPT combination. However, one would run into the insufficient-sample-size

problem frequently when implementing this ideal approach for the lognormal Model. As aforementioned, the average number of cases per CPT combination ranges from 1 to 7.54 among the various departments.

- Departmental Sample-mean Model

This model takes the sample mean of the case durations within a service department in the training set, and treat it as the prediction for the cases within the same department in the test set, namely

$$\widehat{Y}_z^{new} = \frac{1}{n_t} \sum_{i=1}^{n_t} y_i.$$

- Hybrid Sample-mean Model

When departmental sample means are used, the individuality of each surgery, which is manifested by various CPT codes, is lost. Meanwhile, such individuality often results in lack of historical data. Taking into account both concerns, our third benchmark model calculates sample means differently for different surgeries based on the existence of historical data. For a surgery in the test set, if its CPT combination can be found in the training set, the mean of all the surgeries with the same CPT combination is the predicted duration; otherwise, the departmental sample mean serves as the predicted value.

Inclusion of the lognormal model in our comparison is easily understood since previous research has argued for its use. Sample means are intuitive benchmarks because they are common practices when surgery schedules are determined in hospitals. One can certainly find drawbacks in these benchmark models or their implementations. But, the lack of a better benchmark model also validates the necessity of our work. We believe our proposed prediction models in this paper set a reasonable benchmark for future research.

### 4.3 Comparison

Three departments, “EYE”, “OTH”, and “RAD”, have too few surgical cases, which are 88, 65, and 2, respectively, and will be omitted in this section for prediction and comparison. We apply two proposed regression-based methods and three benchmark methods to the remaining 16 departments. When reporting our results, we use “Reg”, “LogReg”, “Lognormal”, “Dept-mean”, and “Hybrid-mean” to represent the linear regression model (1), the log-regression model (2), the lognormal model, the departmental sample-mean model, and the hybrid sample-mean model, respectively. For each department, all the five models are employed to make predictions over the test data sets

Dept	Mean Squared Error				
	Reg	LogReg	Lognormal	Dept-mean	Hybrid-mean
NS	<b>0.828</b> (0.003)	0.851(0.004)	1.865(0.005)	1.848(0.006)	1.086(0.004)
ORT	<b>0.417</b> (0.001)	0.551(0.002)	1.026(0.001)	1.024(0.001)	0.576(0.001)
TPT	<b>0.711</b> (0.003)	0.891(0.003)	1.704(0.005)	1.700(0.005)	0.787(0.004)
URO	<b>0.403</b> (0.001)	0.487(0.001)	1.272(0.002)	1.272(0.002)	0.530(0.001)
CT	0.899(0.003)	<b>0.898</b> (0.003)	2.769(0.004)	2.682(0.005)	1.539(0.005)
THO	0.730(0.004)	<b>0.708</b> (0.004)	0.913(0.004)	0.912(0.004)	0.734(0.004)
UMC	<b>0.095</b> (0.000)	0.186(0.001)	0.377(0.001)	0.375(0.001)	0.107(0.000)
GEN	<b>0.532</b> (0.001)	0.569(0.001)	1.157(0.001)	1.156(0.001)	0.571(0.001)
ONC	<b>0.559</b> (0.001)	0.605(0.001)	1.875(0.002)	1.872(0.002)	1.009(0.002)
GYN	<b>0.371</b> (0.001)	0.424(0.001)	0.885(0.001)	0.884(0.001)	0.490(0.001)
ORA	<b>0.497</b> (0.002)	0.552(0.003)	2.714(0.007)	2.705(0.007)	0.550(0.005)
PLA	<b>0.707</b> (0.003)	2.281(0.060)	2.940(0.005)	2.926(0.005)	1.414(0.004)
VAS	<b>0.540</b> (0.002)	0.588(0.002)	1.349(0.004)	1.348(0.004)	0.799(0.003)
PDS	0.190(0.001)	0.252(0.002)	0.418(0.001)	0.416(0.001)	<b>0.176</b> (0.001)
ENT	<b>0.282</b> (0.002)	0.493(0.003)	1.126(0.002)	1.118(0.003)	0.406(0.002)
POD	0.109(0.001)	0.131(0.001)	0.148(0.001)	0.148(0.001)	<b>0.095</b> (0.001)

Table 6: Mean squared errors of out-of-sample prediction of surgical times for several competing methods. Numbers shown are means and corresponding standard derivations, based on 1000 random splits of the data into training and test sets. Unit: hour.

coming from the 1,000 random splitting of the surgical data (with  $S = \bigcup_{k=1}^4 S_k$ ). Both MSE and MRAE are calculated. Results of the comparison are summarized in Tables 6 and 7.

The highlighted numbers in the two tables represent the smallest MSE or MRAE of prediction, or the best performance, in each respective department. From Tables 6 and 7, we observe the following:

- When MSE/MRAE is used as performance measure, the highlighted numbers occur in the first two columns for 14/15 out of 16 departments. This demonstrates the superior performance of our proposed regression-based methods, as compared to the three benchmark models.
- Between the two of our proposed methods, the linear regression model claims the best performance more often than the log-regression model. This observation suggests that when using CPT codes as explanatory variables to make predictions, the benefit of applying a logarithm transformation to the data is no longer obvious.
- Our proposed regression-based methods significantly improve the two benchmark methods that make predictions based on departmental data. This is consistent with the observation in existing literature that CPT code information plays an important role in predicting surgical case durations.

Dept	Mean Relative Absolute Error				
	Reg	LogReg	Lognormal	Dept-mean	Hybrid-mean
NS	0.440(0.001)	<b>0.402</b> (0.001)	1.313(0.003)	1.238(0.003)	0.608(0.001)
ORT	<b>0.377</b> (0.000)	0.403(0.000)	0.865(0.001)	0.844(0.001)	0.489(0.001)
TPT	<b>0.356</b> (0.001)	0.461(0.001)	0.995(0.002)	0.959(0.002)	0.390(0.001)
URO	<b>0.474</b> (0.001)	0.568(0.001)	1.277(0.001)	1.266(0.001)	0.561(0.001)
CT	0.275(0.001)	<b>0.253</b> (0.001)	1.221(0.003)	1.113(0.003)	0.570(0.001)
THO	<b>0.406</b> (0.001)	0.452(0.003)	0.680(0.004)	0.667(0.004)	0.463(0.002)
UMC	<b>0.332</b> (0.000)	0.523(0.001)	1.789(0.002)	1.684(0.002)	0.408(0.001)
GEN	<b>0.387</b> (0.000)	0.406(0.000)	0.943(0.001)	0.922(0.001)	0.461(0.000)
ONC	<b>0.354</b> (0.000)	0.396(0.000)	1.130(0.001)	1.091(0.001)	0.585(0.001)
GYN	<b>0.339</b> (0.000)	0.436(0.001)	0.962(0.001)	0.930(0.001)	0.515(0.001)
ORA	0.502(0.002)	<b>0.477</b> (0.002)	1.445(0.005)	1.395(0.005)	0.529(0.003)
PLA	<b>0.401</b> (0.001)	0.453(0.001)	1.881(0.002)	1.745(0.002)	0.802(0.003)
VAS	0.305(0.001)	<b>0.304</b> (0.001)	0.608(0.001)	0.602(0.002)	0.378(0.001)
PDS	<b>0.514</b> (0.002)	0.717(0.002)	1.592(0.003)	1.506(0.003)	0.536(0.001)
ENT	<b>0.491</b> (0.001)	0.831(0.002)	2.975(0.005)	2.747(0.004)	0.956(0.003)
POD	0.420(0.001)	0.466(0.002)	0.549(0.002)	0.546(0.002)	<b>0.386</b> (0.001)

Table 7: Mean relative absolute errors of out-of-sample prediction of surgical times for several competing methods. Numbers shown are means and corresponding standard derivations, based on 1000 random splits of the data into training and test sets. Unit: hour.

- The hybrid sample-mean method, which utilizes the CPT code information, does perform better than the other two benchmark methods, which do not use the CPT code information. The hybrid sample-mean method cannot out-perform our regression-based methods for most departments, because it uses the CPT code information only when there is an exact match of CPT combination in the training set. Understandably, the hybrid sample-mean method performs well only when there are “sufficient” number of historical cases with the same CPT combination. In practice, it is not always easy to decide how many are enough, and there do exist circumstances when there are only a handful of historical cases or there is no such case at all. Looking at the comparison result tables, there are several departments (e.g., CT, PLA, ENT) where the hybrid-mean predictions have MRAEs almost as twice large as those using our regression methods. Similar large differences in MSE can also be found. This observation suggests that the hybrid sample-mean method is not a suitable tool for predicting surgical durations when numerous and complex CPT combinations are used.
- For 8 of the 16 departments, the reduction of MRAE by using the linear regression model instead of the hybrid sample-mean method (the best performer of the three benchmark methods) is bigger than 0.10, which corresponds to a 30-minute reduction of prediction error for a 5-hour long surgery.

- The two benchmark models, the lognormal model and the departmental sample-mean model, have similar performances. The lognormal model does not exhibit any noticeable edge in terms of prediction quality over the simple sample-mean model. To some extent, this result “validates” the use of the sample-mean model in practice. We believe that the lack of difference between these two benchmark models is due to the fact that the surgical data within a department do not always follow a lognormal distribution (see Section 3 and Figure 1).

## 5 Conclusion

This paper presents regression-based methodologies that take multiple CPT codes as explanatory variables when predicting surgical case durations. Our research is motivated by the fact that CPT codes describe how a surgical case should be performed, and thus provide specific knowledge and information relevant to individual surgical cases. The importance of CPT codes in predicting surgical case durations has been noted in health care literature for years. Our research demonstrates the benefit of utilizing CPT code information by a prediction comparison using real data from a large central Texas hospital. The reduction of prediction errors due to efficient utilization of CPT codes will certainly boost certainty in the scheduling process and help cut “white spaces” between surgeries (i.e., buffer time inserted between surgeries to accommodate variability) or overruns so that more surgeries can be scheduled with a higher start time reliability. Our proposed methodology could help a great deal with the issues related to operating room scheduling and resource utilization, and consequently, will bring considerable economic benefits to the bottom line of a hospital and lead to greater patient satisfaction.

To the best of our knowledge, our paper is the first that predicts surgery case durations based on multiple CPT codes that a surgical case performs. In our opinion, one of the reasons that such a predictive model was not available prior to our research is perhaps caused by the complexity involved in devising a proper design matrix. If naively constructing a design matrix according to the appearances of CPT codes in surgical cases, one will likely end up with an ill-conditioned matrix that is not solvable. In our research, we develop general procedures to overcome this difficulty by systematically grouping CPT combinations and treat not only single CPT codes but also distinctive CPT combinations with multiple CPT codes as separate explanatory variables. Our algorithm guarantees a fully ranked design matrix, and consequently, the solvability of the least-squares estimation. Although the implementation details are provided for surgical cases using up to eight CPT codes (which in itself has already represented very complicated surgeries), our

models and algorithms can be applied to surgical cases using any number of CPT codes or any combination of CPT codes.

One possible extension of our research is to consider other important covariates together with CPT codes. In addition to CPT codes, which is arguably the most important factor relevant to the prediction of surgical case durations, prior research also identified a number of other factors influencing surgical case durations (such as surgeon experience, anesthesia type, patient’s status). The inclusion of those factors is methodologically straightforward — we can extend our predictive models by simply adding the relevant covariates. Although we believe that an extended model that incorporates both CPT codes and other relevant covariates as explanatory variables has the potential to further reduce prediction uncertainty, testing the extended model using real data would require a different data set than the one we have, and another round of (possibly very lengthy) data collection efforts.

## Acknowledgement

Huang’s work was partially supported by grants from the National Science Foundation (DMS-0606580) and the National Cancer Institute (CA57030).

## References

- Casella, G. and R. L. Berger (2001). *Statistical Inference, 2nd Edition*. Duxbury.
- Combes, C., N. Meskens, C. Rivat, and J. Vandamme (2008). Using a KDD process to forecast the duration of surgery. *International Journal of Production Economics* 112, 279–293.
- Dexter, F., E. Dexter, D. Masursky, and N. Nussmeier (2008). Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesthesia and Analgesia* 106, 1232–1241.
- Litvak, E. and M. Long (2000). Cost and quality under managed care: Irreconcilable differences? *American Journal of Managed Care* 6, 305–312.
- May, J., D. Strum, and L. Vargas (2000). Fitting the lognormal distribution to surgical procedure times. *Decision Sciences* 31, 129–148.
- McManus, M., M. Long, A. Cooper, J. Mandell, D. Berwick, M. Pagano, and E. Litvak (2003). Variability in surgical caseload and access to intensive care services. *Anesthesiology* 98, 1491–1496.

- Mitchell, T. (1997). *Machine Learning*. New York: McGraw-Hill.
- Olivares, M., C. Terwiesch, and L. Cassorla (2008). Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Science* 54, 41–55.
- Spangler, W., D. Strum, L. Vargas, and J. May (2004). Estimating procedure times for surgeries by determining location parameters for the lognormal model. *Health Care Management Sciences* 7, 97–104.
- Strum, D., J. May, A. Sampson, L. Vargas, and W. Spangler (2003). Estimating times of surgeries with two component procedures. *Anesthesiology* 98, 232–240.
- Strum, D., J. May, and L. Vargas (2000). Modeling the uncertainty of surgical procedure times: comparison of the log-normal and normal models. *Anesthesiology* 92, 1160–1167.
- Strum, D., A. Sampson, J. May, and L. Vargas (2000). Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology* 92, 1454–1466.
- Weisberg, S. (2005). *Applied Linear Regression, 3rd Edition*. Wiley/Interscience.
- Weiss, E. (1990). Models for determining estimated start times and case ordering in hospital operating rooms. *IIE Transactions* 22, 143–150.
- Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition*. Morgan Kaufmann.
- Wright, I., C. Kooperberg, B. Bonar, and G. Bashein (1996). Statistical modeling to predict elective surgery time: Comparison with a computer scheduling system and surgeon-provided estimates. *Anesthesiology* 85, 1235–1245.