Length of Stay: Prediction and Explanation

by David H. Gustafson

Five methodologies for predicting hospital length of stay were developed and compared. Two—a subjective Bayesian forecaster and a regression forecaster—also measured the relative importance of the symptomatic and demographic factors in predicting length of stay. The performance of the methodologies was evaluated with several criteria of effectiveness and one of cost. The results should provide encouragement for those interested in computer applications to utilization review and to scheduling inpatient admissions.

Advances in medical diagnosis and treatment have placed a heavy burden on our medical care system. The public, while demanding the benefits of these improved services, is simultaneously unwilling to pay for the concurrent cost increases. In the resulting efforts to find better ways to allocate those facility, equipment, and manpower resources necessary for hospital operation, techniques have been developed for scheduling elective admissions, predicting bed needs, and measuring bed utilization. One component in these techniques is an accurate prediction of how long a patient will stay in the hospital and an understanding of the factors that influence his stay.

Alternative Approaches

This report describes the development, demonstration, and comparison of five methodologies for predicting and, in two cases, explaining hospital length of stay. Of these methodologies, three gave a point estimate of the length of stay, and were based on physicians' subjective opinions, while two gave a probability distribution over all lengths of stay, and were based on empirical data.

Empirical predictors can use criteria such as minimum least squares deviation, maximum likelihood estimation, or maximum posterior probability [1]. All are effective when a sound empirical data base can be collected. However, if the necessary data are unavailable (frequently the case in length-of-stay estimation), or if the data-generating process is unstable, empirical parameter estimation will suffer. The varying degrees of success that Bartscht [2] and Robinson [3] found when testing multiple linear regression length-of-stay predictors might be attributed to such variations in data quality.

Similarly, subjective length-of-stay estimates have met with varying degrees of success [4,5]. Although they do not require a massive collection of frequently inaccessible data, they do draw upon the subjective judgments of experts. Normally, estimators do not accurately extract or combine the impacts of all the information inherent in data [6]. However, as will be shown, recent advances in the areas of Bayesian statistics and human information-processing have markedly improved subjective prediction of uncertain events.

A forecast may be given in terms of either a point estimate or a distribution estimate. The point estimate, common in length-of-stay forecasts, is easier to develop and is useful in predicting events of high certainty; however, it ignores much information inherent in processes of a highly stochastic nature. For distribution estimates, two processes available are the Direct and the Bayesian. The former directly estimates the probability of a length of stay of i days for a patient defined by his demographic characteristics, $D_1, \ldots, D_k$, and symptomatic characteristics, $D_{k+1}, \ldots, D_n$. Let H_i denote the event that the patient stays i days. The posterior probability estimate will then be:

$$P(H_i \mid D_1, \ldots, D_k, D_{k+1}, \ldots D_n)$$

Such a technique has characteristics, to be described, that may lead to inaccurate probability estimates.

Alternately, Bayes' Theorem can be used to derive the posterior probability distribution by modifying the prior probability distribution of length of stay, $P(H_i)$, in the light of demographic and symptomatic patient data expressed in terms of the likelihood function $P(D_1, \ldots, D_k, \ldots, D_n \mid H_i)$. Bayes' Theorem can be expressed as:

$$P(H_i \mid D_1, \ldots, D_k, \ldots, D_n) = \frac{P(D_1, \ldots, D_k, \ldots, D_n \mid H_i)\ P(H_i)}{\sum_i P(D_1, \ldots, D_k, \ldots, D_n \mid H_i)\ P(H_i)} \quad (1)$$

Distribution estimates, like point estimates, may be obtained either empirically or subjectively. Both the Bayesian and the Direct subjective distribution techniques employ personal estimates of probabilities, defined as a person's *degree of belief,* based on certain evidence, in a given event. In simple terms, a posterior probability estimate of an event could be described as the amount the estimator is willing to bet, against $1 put up by the house, that the event will occur.

Several studies have shown that personal probability estimators tend to be conservative. They typically underestimate the impact of a datum (e.g., blood pressure = 200/120) on an hypothesis (e.g., length of stay = 8 days) [6-10]. It may be that estimators hesitate to estimate values close to the 0.0 and 1.0 probability boundaries, for when odds are substituted for these boundaries conservatism is reduced [11-13]. Probabilities can be retrieved from odds as follows:

$$P(H_i \mid D) = \left\{ \sum_j \left[\frac{P(H_j \mid D)}{P(H_i \mid D)} \right] \right\}^{-1} \quad (2)$$

Odds also permit a more sensitive differentiation between small probabilities. Suppose a datum has a low likelihood for either of two hypotheses:

$$P(D \mid H_1) = 0.0009$$
$$P(D \mid H_2) = 0.00009$$

Both values are so small as to be difficult to estimate accurately, but a reasonable estimate of their ratios would be much easier:

$$\frac{P(D \mid H_1)}{P(D \mid H_2)} = \frac{10}{1}$$

Conservatism also appears to be inversely related to the estimator's understanding of the data-generating process. Hence, conservatism is reduced by selecting experts in the project's field of interest as the estimators [6, 14]. One would expect, for instance, that a physician, rather than a lawyer, would better predict length of stay because he better understands how diagnosis, blood pressure, and other patient data affect length of stay.

Finally, the more data the estimators must combine, the greater their conservatism appears to be [6]. That is, symptoms and vital signs may tend to overload, rather than help, the physician estimator. This may be a primary cause of conservatism in direct posterior probability estimates, where the physician is given a vast amount of patient data, $D_1, \ldots, D_n$, and asked to assimilate and process it in order to give the posterior probability estimate $P(H_i \mid D_1, \ldots, D_n)$. Bayes' Theorem, in the form of Equation 1, requires the physician to estimate a likelihood $P(D_1, \ldots, D_n \mid H_i)$. If n is large, this task would also be nearly impossible to do well. However, the information overload can be eliminated if the data are conditionally independent. The physician can consider one datum at a time, $P(D_j \mid H_i)$, and Bayes' Theorem can combine the data. $D_1, \ldots, D_n$ are conditionally independent given $H = H_i$ if $P(D_1, \ldots, D_n \mid H_i) = P(D_1 \mid H_i) \ldots P(D_n \mid H_i)$ for all i; thus, the equation can be written as follows:

$$P(H_i \mid D_1, \ldots, D_n) = \frac{P(D_1 \mid H_i) \ldots P(D_n \mid H_i)\ P(H_i)}{\sum_i P(D_1 \mid H_i) \ldots P(D_n \mid H_i)\ P(H_i)} \qquad (3)$$

The Bayesian and Direct methods of estimating personal probability distributions have been compared in controlled environments by several researchers [14-16]. Generally, the results indicate that the Bayesian method is superior to the Direct, but when all data have a strong influence on the problem the difference between the two techniques tends to disappear. Because the relative performance of these two techniques is still open to question, both were investigated in this experiment.

Information Levels

The data used were from a sample of eight inguinal herniotomy patients, stratified over four categories of length of stay, selected from the 39 such patients admitted to the fourth surgical division of the Henry Ford Hospital, Detroit, during January and February of 1966. Five methods for predicting length of stay were compared.

In order to investigate the relative predictability of the methodologies, and to assess the impact of demographic information on length-of-stay predictors, four levels of information were used:

1. Symptomatic (but not demographic) information available after admission and preliminary examination
2. Symptomatic and demographic data available after admission and preliminary examination
3. Symptomatic and demographic data available directly after herniotomy
4. Symptomatic and demographic data available three days after herniotomy (if there are to be complications, most should have occurred by this time)

Methodologies

Point Estimators

Method 1. Subjective Point Estimates—Subjective point estimates were obtained from three physician groups: (1) three nonattending surgical residents; (2) three nonattending board-certified surgeons; and (3) one of the patient's attending surgeons. Each physician had to select his prediction from one of twelve hypothesized lengths of stay:

$H_1 = 5$ days or less
$H_2 = 6$ days
$H_3 = 7$ days
$H_4 = 8$ days
$H_5 = 9$ days
$H_6 = 10$ days
$H_7 = 11$ days
$H_8 = 12$ days
$H_9 = 13$ days
$H_{10} = 14$ days
$H_{11} = 15$ days
$H_{12} = 16$ days or more

The hypotheses were truncated at 5 and 16 days because 98 per cent of all lengths of stay for hernia patients were within this range. The first two physician groups were given abstracts of the patient's medical record and asked to estimate length of stay. The attending surgeon's estimates, based on his direct experience, were obtained by interview while the patient was in the hospital. Since it was impossible to control the amount of information available to the attending surgeon, the impact of demographic data on length of stay was not subject to rigorous evaluation; however, the attending surgeon did not believe that these data influenced his predictions.

Method 2. Regression Analysis—The second methodology involved multiple linear regression analysis. The empirical data source was a computer system for storing abstracts of medical records (these abstracts will be referred to as "profiles"), which produced 188 useful hernia records [17]. The factors in the theoretical model were obtained (1) by having a surgeon rate, on a discrete scale from 0 to 5, the importance, for predicting length of stay, of factors included in the profile; (2) by searching the literature for factors that might influence length of stay; and (3), through discussions with three surgeons, by modifying the model to include nonlinear and interaction components.

The final model included linear, binary, logarithmic, and interaction terms. The coefficients of this theoretical model were estimated by the "step-wise" procedure [18], which sequentially generates the regression, entering variables in order of their relative importance until all significant variables are included. This procedure replaces the original model with a smaller one based entirely upon the data available. However, in using this method the function of least squares estimation was solely to estimate coefficients, not to change the model, so coefficients were estimated for all variables that entered above an F level of 0.001. Table 1 shows the theoretical model for each information level.

Method 3. Historical Mean—The third methodology involved the determination of the average length of stay for all herniotomy patients discharged from Henry Ford Hospital during 1965.

Distribution Estimators

Method 4. Direct Posterior Odds Estimation—The fourth methodology was a direct method of estimating a subjective probability distribution, using odds rather than probabilities to reduce conservatism. It employed three surgical residents, who, on the basis of available information, selected the most likely length of stay, H_B, and estimated:

$$\frac{P(H_B \mid D_1,\ldots,D_n)}{P(H_i \mid D_1,\ldots,D_n)},\ i = 1,\ldots,12.$$

The $P(H_i \mid D_1,\ldots,D_n)$, $i = 1,\ldots,12$ were then retrieved from these odds via Equation 2.

Method 5. Bayes' Theorem—The fifth methodology, employing Bayes' Theorem to combine the impacts of the data complexes on the hypothesized lengths of stay, had three variations. The first employed subjective likelihood estimates of data classified into conditionally independent complexes. The second used data arbitrarily classified into complexes. The third used a combination of subjective and actuarial likelihood estimates of the conditionally independent complexes.

Discussions with surgeons showed that there was strong conditional dependence between some of the data. To retain the aggregation benefits of Bayes' Theorem, the data were placed into sets or complexes, C_k, within which the data were highly dependent but between which little conditional dependency existed. In order to further reduce conservatism, the likelihoods were estimated in terms of odds rather than probabilities:

$$\frac{P(C_k \mid H_B)}{P(C_k \mid H_i)},\ k = 1,\ldots,\ 13, \text{ and } i = 1,\ldots,12.$$

Table 1. Descriptions of the theoretical regression model at the four information levels

Item of Information	*Information Level**				*Relation*			
	1	*2*	*3*	*4*	*Linear*	*Binary*	*Logarithmic*	*Multiplicative*
Day of admission		x	x	x		x		
Referral classification		x	x	x		x		
Method of payment		x	x	x		x		
Marital status		x	x	x		x		
Occupational activity		x	x	x		x		
Type of employment		x	x	x		x		
Recurrence status	x	x	x	x		x		
Age of patient		x	x	x		x		
Sex		x	x	x		x		
Race		x	x	x		x		
Type of hernia	x	x	x	x		x		
Type of anesthesia			x	x		x		
Number of:								
Diagnostic procedures (preoperative)	x	x	x	x	x		x	x
Consultations (preoperative)	x	x	x	x	x		x	x
Lab tests (preoperative)	x	x	x	x	x		x	x
Diagnostic procedures (postoperative)				x	x		x	x
Lab tests (postoperative)				x	x		x	x
Consultations (postoperative)				x	x		x	x
Drugs being taken	x	x	x	x	x			
Medical problems	x	x	x	x	x			
Positive findings	x	x	x	x	x			
Contents of hernia sac			x	x	x			
Operative procedures			x	x	x			
Postoperative complications				x	x			
Preoperative length of stay			x	x	x			

*Levels 1 and 2, preoperative—Level 1, symptomatic data only; Level 2, all data. Level 3, immediately postoperative; Level 4, three days postoperative.

This necessitated the use of the ratio form of Bayes' Theorem, combining the separate conditional probabilities for H_B and H_i as follows:

$$\frac{P(H_B \mid C_1, \ldots, C_m)}{P(H_i \mid C_1, \ldots, C_m)} = \frac{P(C_1 \mid H_B)}{P(C_1 \mid H_i)} \cdots \frac{P(C_m \mid H_B)}{P(C_m \mid H_i)} \; \frac{P(H_B)}{P(H_i)} \,. \qquad (4)$$

When the data source is limited, there is no statistically sound yet operationally feasible method for pooling data $D_1, \ldots, D_n$ into conditionally independent complexes $C_1, \ldots, C_m$. An empirical test of conditional independence is only asymptotically valid, so the amount of data required to test even first order conditional independence is frequently prohibitively large. We say that $D_1, \ldots, D_m$ are marginally independent when

$$P(D_1, \ldots, D_m) = P(D_1) \ldots P(D_m).$$

If marginal independence implied conditional independence, previously defined, the sample might be reduced to a tractable size. However, many examples show that this is not the case [19]. Fortunately, the probability obtained by Bayes' Theorem is not strongly influenced by weak dependencies [20], so that higher order interactions, important in more rigorous tests of conditional independence, can be ignored.

Three steps were taken to place data into complexes of low conditional interdependency:

1. A surgeon reviewed the profile factors to determine which must be pooled when predicting length of stay.
2. Three surgeons, six surgical residents, and two interns sorted the same factors, recorded on 3 x 5 cards, into sets having high intradependency and low interdependency. Whenever three or more physicians identified certain factors as being conditionally dependent, they were so classified.
3. Results were compared; disagreements were resolved by the chief of the surgical division. After seeing a data complex, C_k, the physician selected the hypothesized length of stay, H_B, from which the data had most probably come, and estimated how much more likely it was that the data had come from H_B than from any other H_i:

$$\frac{P(C_k \mid H_B)}{P(C_k \mid H_i)} \text{ for } i = 1, \ldots, 12.$$

To test the method of selecting data complexes with low conditional dependency, the Bayesian predictions from these likelihood ratio estimates were compared with a second variation: Bayesian predictions using randomly combined, and so presumably conditionally dependent, data complexes. Although the same residents gave both estimates, transfer-of-learning effects were removed by appropriate randomization in data presentation.

The third variation of Bayes' Theorem also used likelihood ratios, but the impacts from four of the "conditionally independent" complexes were estimated empirically. Unfortunately the data source (1100 observations) was insufficient to give reliable estimates of $P(C_k \mid H_i)$ *for* $i = 1, \ldots, 12$ and C_k having as many as 10 levels. So the results of this methodology must be viewed critically.

Figure 1 summarizes the experimental design and shows the abbreviations used for the various methodologies.

Training

The nine physicians who described their uncertainty about length of stay in terms of odds or likelihood ratios were given two hours of the following

Methods (and Abbreviations Used)	*Number of Physician Estimators*	
Subjective Point Estimates:		All estimates for eight patients at four information levels*
Resident (P.R.)	3	
Surgeon (P.S.)	3**	
Attending Surgeon (A.S.)	1	
Multiple Linear Regression Analysis (RGN)	None	
Historical Mean (M)	None	
Direct Posterior Odds Estimation		
Resident (P.O.)	3	
Bayesian:		
Subjective, Conditionally Independent (B.I.)	3	
Subjective, Conditionally Dependent (B.D.)	3	
Hybrid, Conditionally Independent (B.H.)	3	

***Levels of Information**

Level 1	*Level 2*	*Level 3*	*Level 4*
Preoperative	Preoperative	Immediately Postoperative	Three Days Postoperative
Symptomatic data only	All data	All data	All data

**Although the design included three nonattending surgeons, one was unable to complete his estimates in time for the analysis.

Fig. 1. Summary of Experimental Design

training. The doctor was asked to imagine two book bags filled with red and blue poker chips in a specified proportion, but with the higher proportion being red chips in Bag R and blue chips in Bag B. One bag was selected and chips were drawn randomly from it, one at a time, with replacement. Actually, the sequence of draws was programmed ahead of time and transmitted to the physician via rows of red and blue lights on a display board. After each draw the doctor specified which bag was more likely to have been chosen, and how much more likely.

Here Bayes' Theorem is the normative model to which we can compare the doctors' behavior. If p represents the proportion of red chips in Bag R and of blue chips in Bag B, then $1-p$ represents the proportion of blue chips in Bag R and of red chips in Bag B. Thus, for Bag B the probability of getting s "successes" (blue chips) in n draws is proportional to $p^s\,(1-p)^s$. Therefore, the likelihood ratio of the s successes in n draws for Bag B versus Bag R is

$$L = \frac{P(s,n \mid B)}{P(s,n \mid R)} = \frac{p^s(1-p)^{n-s}}{p^{n-s}(1-p)^s} = \left(\frac{p}{1-p}\right)^{2s-n}$$

Since prior opinion was uniformly distributed (each bag was equally likely to be chosen), the posterior odds of Bag B versus Bag R, given *s* and *n*, is

$$\left(\frac{p}{1-p}\right)^{2s-n} .$$

This approach permitted calculation and comparison of correct odds with the estimates. Feedback was given to the physicians.

Rather than being fixed throughout the training exercise, the proportions of chips were alternated among 70-30, 55-45, and 85-15. This presented the physician with conditions of varying uncertainty. A single draw from a bag containing chips in 85-15 proportions has much more diagnostic value than does a single draw from a 55-45 bag. His odds estimation should similarly reflect this difference in uncertainty. The 70-30 bag was used to portray a condition of uncertainty somewhere between the other two, thus acquainting the doctor with numerical description at several levels of uncertainty.

Results

Measures of Effectiveness

Because some forecasting techniques gave distribution estimates and others gave point estimates, it was difficult to measure their relative effectiveness. Two parameters (the mode and the mean of the distribution) were compared, one at a time, with the point estimates.

The deviation from the mode, D_m, compared the patient's actual length of stay with the point estimate or with the mode of the distribution estimate:

$$D_{mi} = | X_{ai} - \hat{X}_{mi} | ,$$

where X_{ai} = actual length of stay for patient *i*.

$$\hat{X}_{mi} = \begin{cases} \text{mode (if a distribution estimate)} \\ \text{point estimate (otherwise)} \end{cases}$$

The second measure, the deviation from the mean, D_a, compared the actual length of stay with the point estimate or with the mean of the distribution estimate:

$$D_{ai} = | X_{ai} - \hat{\bar{X}}_{mi} | ,$$

$$\text{where } \hat{\bar{X}}_{mi} = \begin{cases} \text{mean (if a distribution estimate)} \\ \text{point estimate (otherwise)} \end{cases}$$

A third measure, M_c, the number of times that D_m was equal to 0, was also used to indicate whether or not the prediction was correct. This measure can be valuable in selecting a technique in cases when a prediction that is only close to being correct will not be good enough.

Figure 2 (next page) presents the frequency distribution of D_m values for each estimation technique. A negative value is an underestimation; a positive value, an overestimation. Only the subjective Bayesian techniques, (B.I.) and (B.D.), have a zero mode. It is interesting to note that these are the only subjective techniques that do not require data aggregation by physicians, and that also do not directly request an estimate of length of stay. Rather, they ask for information about the data, that is,

$$\frac{P(C_k \mid H_B)}{P(C_k \mid H_i)} \cdot$$

It may be that this process removes a tendency on the part of the physician to overestimate length of stay.

The data points clustered between −7 and −9 were primarily the result of one case, in which a patient, three days after the operation, developed complications that added nine days to his length of stay. Such differences between patients have been accounted for in the analysis of variance.

Figures 3, 4, and 5 (pages 23-25) illustrate performance techniques of information at each level, and as an overall average of information levels, in terms of the three measures just described. Each point represents an average for all doctors' estimates and for all patients. The data indicate that the historical average is the poorest predictor, with regression analysis next (Figure 3). However, at the "operative" level regression analysis does quite well, according to the D_m and D_a measures. This may be explained by the entrance of a variable with high diagnostic value, the preoperative length of stay. Increasing the amount of data does not appear to improve predictions by the P.O. or P.R. techniques, as is the case for the other techniques; in fact, their relative performance, initially high, worsens as data are added. This may be a result of the aggregation difficulties encountered by the estimator.

These comments do not apply to the point estimates by the surgeons (P.S. or A.S.), possibly because of their greater knowledge of the data-generating process (see Figure 4). The attending surgeon, with his broader knowledge of the patient, appears especially adept at handling increased patient data; this is indicated by his strong performance at the postoperative information level.

When the assumption of conditional independence is ignored (B.D.), the performance of the likelihood ratio technique deteriorates (Figure 5). This is especially true at the operative and postoperative information levels, where the deflation or inflation caused by dependencies is more pronounced.

Analyses of variance, used to test the significance of selected portions of the D_a and D_m data, employed a four-factor, partially hierarchical model having subject groups of unequal size nested within the treatment factor [21].

(text continued on page 26)

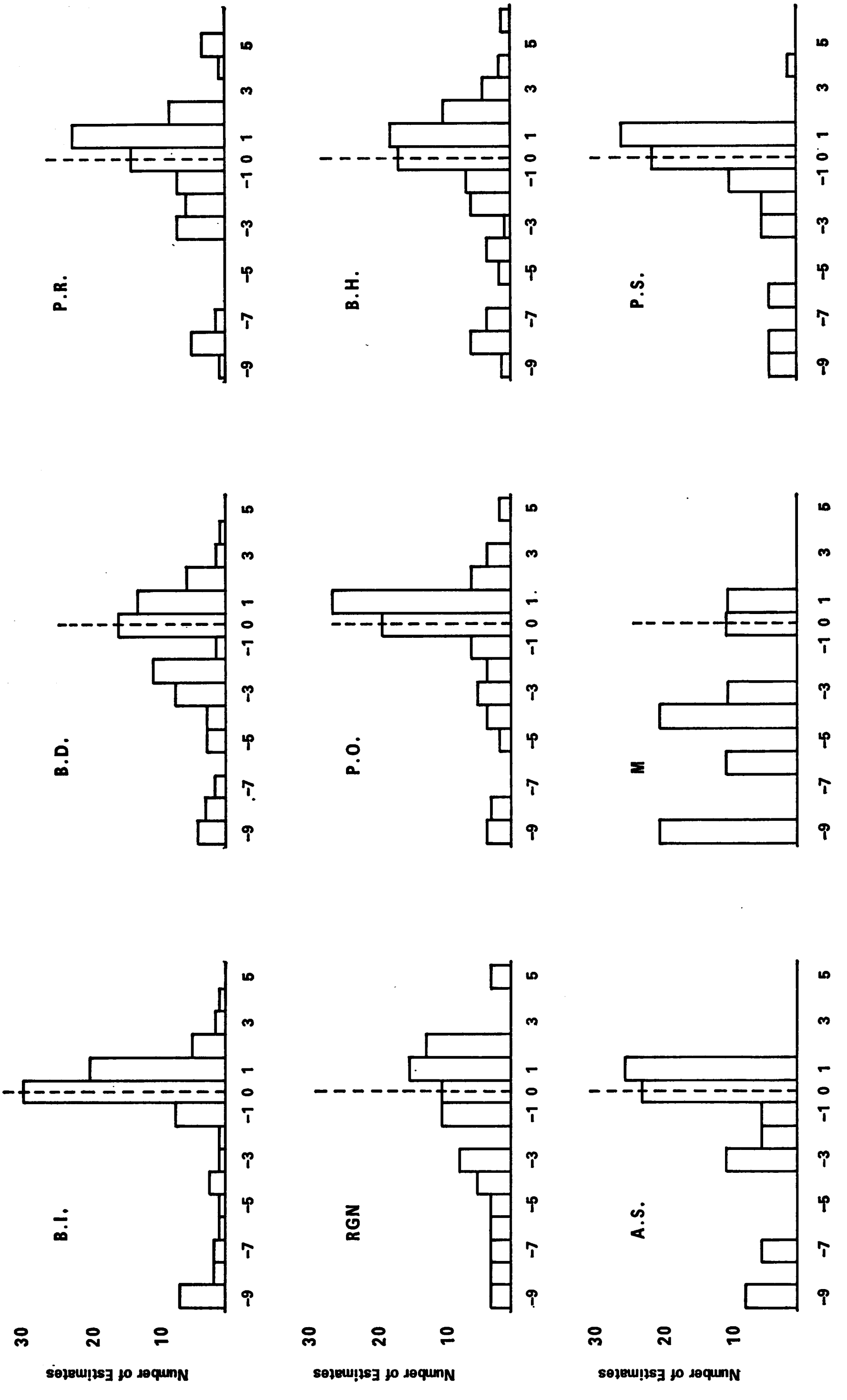


Fig. 2. Distributions of D_m, the deviation from the mode, for all estimation techniques.

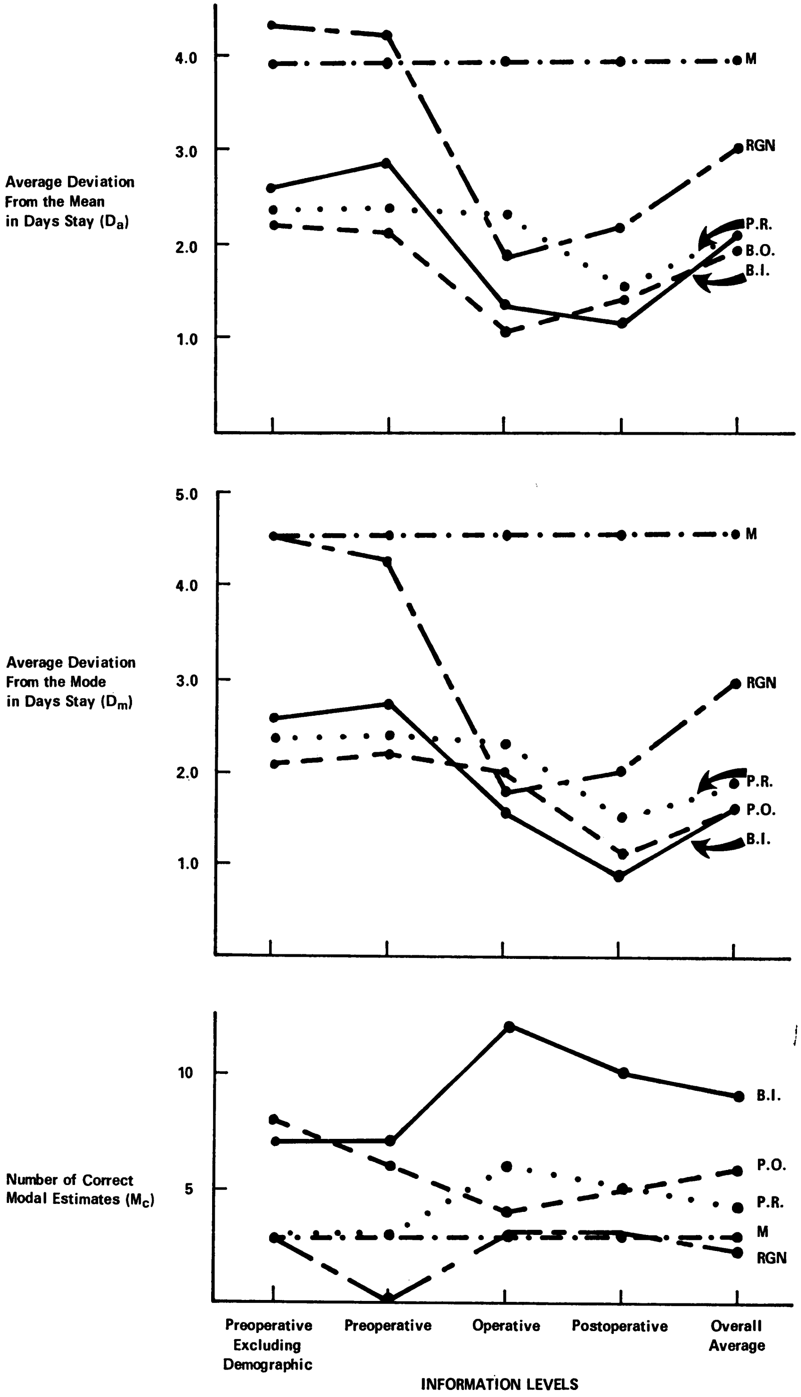


Fig. 3. Comparison of results in the various methodologies with *M*, the historical average prediction, under several information levels.

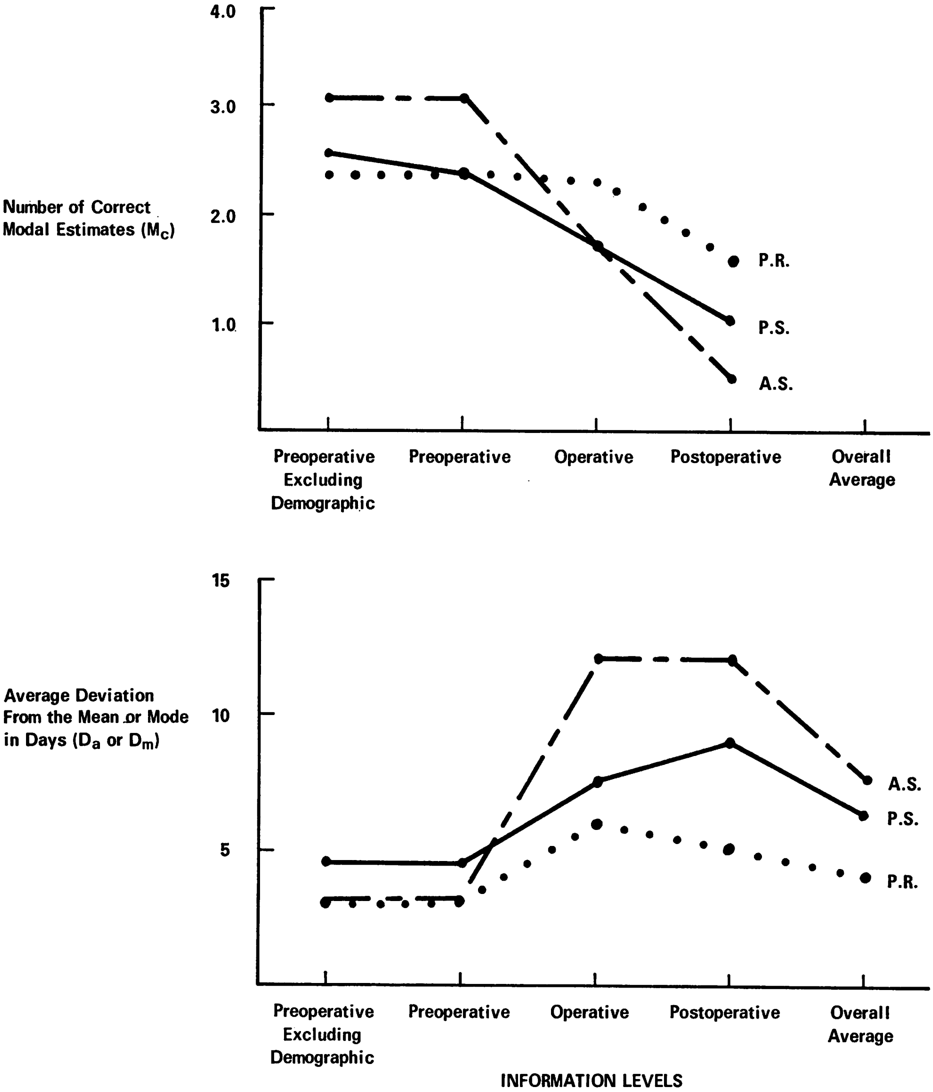


Fig. 4. Comparison of subjective point estimates for physicians with different amounts of medical training (P.R. vs P.S.) and different patient knowledge (P.S. vs A.S.)

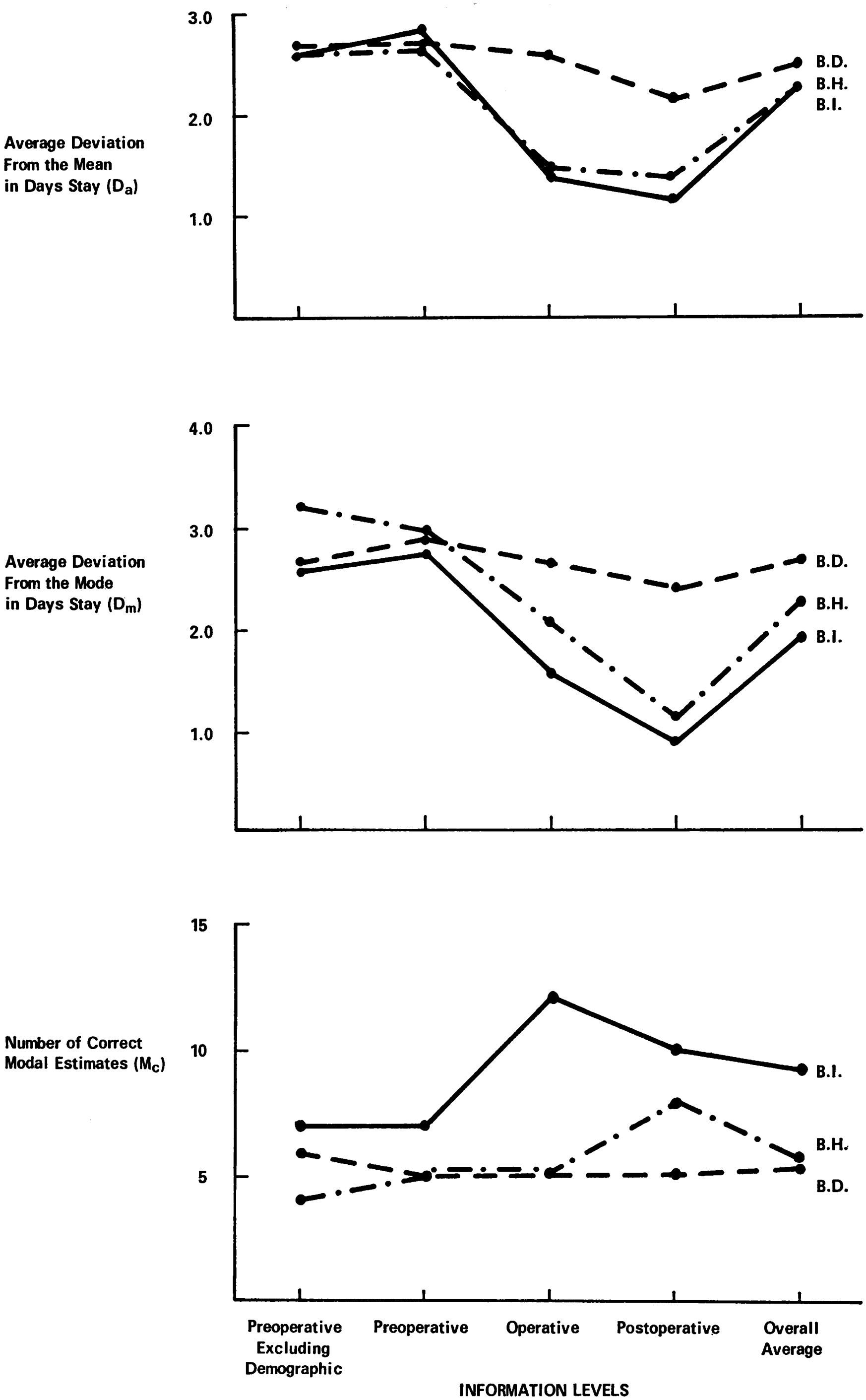


Fig. 5. Comparison of the predictive accuracy of three Bayesian methodologies.

Significant differences were found between prediction techniques and between information levels (Table 2). Scheffe's multiple comparison tests [22] indicated that there were significant differences between (1) the historical mean (M) and all other techniques; (2) the regression analysis (RGN) and the Independent Bayesian (B.I.), Posterior Odds (P.O.), and Resident Point Estimate (P.R.) techniques; (3) the dependent (B.D.) and the conditionally independent (B.I.) Bayesian techniques; and (4) the preoperative information levels and the operative and postoperative information levels (see Table 3).

The significance of the M_c data was investigated by testing two hypotheses: (1) that information level has no significant effect on technique performance and (2) that there is no difference in technique performance. Since the number of correct modal estimates (S_n) comes from a binomial population, a confidence interval could be estimated by using the normal approximation of the binomial distribution. If S_n fell outside this interval, the hypothesis could be rejected. Hypothesis 1 could not be rejected at the .95 level (Table 4), so all information levels were pooled to gain additional sensitivity for testing Hypothesis 2. This hypothesis was rejected at the .95 level, indicating that the conditionally independent Bayesian model (B.I.) was superior, and the regression analysis (RGN) and historical mean (M) were inferior, to all other techniques (Table 5).

Further tests [23], conducted on four of the methodologies, indicated that (1) the results would be duplicated in a closely controlled experimental environment; (2) the Bayesian model's superiority is more evident when performances are compared with normative distributions of uncertainty rather

Table 2. Analyses of variance

Source	*Degrees of Freedom*	*D_a Measure*		*D_m Measure*	
		MS D_a	*F*	*MS D_m*	*F*
Between Subjects					
A—Predictive Methods	8	23.05	5.56[a]	25.82	7.12[b]
B—Subjects within Groups	11	4.04		3.63	
Within Subjects					
C—Patients	7	19.57		140.17	
A × C	56	0.90		3.70	
C × B	77	0.17		4.27	
D—Information	3	549.00	72.31[b]	62.87	16.08[b]
A × D	24	10.00	1.31	2.35	
D × B	33	7.59		3.91	
C × D	21	18.29		37.08	
A × C × D	168	0.99		1.71	
C × D × B	231	0.69		2.16	

[a] Significant at .95
[b] Significant at .99

Table 3. Results of paired comparisons between estimation techniques (given in terms of the level of significance for rejecting the hypothesis of no differences)

Technique	B.D.		RGN		M	
	D_m	D_a	D_m	D_a	D_m	D_a
B.I.	.90	—	.88	—	.99	.82
P.O.	—	—	.92	.82	.99	.90
P.R.	—	—	.75	.75	.99	.90
RGN	—	—	—	—	.75	—
B.H.	—	—	—	—	.95	.82
B.D.	—	—	—	—	.90	.75
P.S.	—	—	—	—	.99	.90
A.S.	—	—	—	—	.95	.75

Table 4. Confidence interval calculation for the hypothesis that information level had no effect on the number of correct modal estimates

Information Level	S_n	n	$P = \frac{S_n}{n}$
Preoperative (symptomatic)	41.5	216	.192
Preoperative (all data)	36.5	216	.169
Operative	57.5	216	.266
Postoperative	60.0	216	.278
Overall	195.5	864	.226

Confidence interval $= 36.40 \leq X \leq 61.12$

Table 5. Confidence interval calculation for the hypothesis that the estimation technique had no effect on the number of correct modal estimates

Estimation Technique	S_n	n	$P = \frac{S_n}{n}$
Event (E)	17.0	96	.177
Posterior Odds (P.O.)	23.0	96	.239
Likelihood Ratio	36.0	96	.375
Likelihood Nonindependent	21.0	96	.218
Likelihood Empirical	22.0	96	.229
Regression (RGN)	9.0	96	.094
Event-Surgeon	25.5	96	.225
Attending Surgeon (A.S.)	30.0	96	.313
Mean (M)	12.0	96	.125
Overall	195.5	864	.226

Confidence interval $= 13.5 \leq X \leq 30.9$

than with the actual length of stay; and (3) a Bayesian model with all data verified to be conditionally independent at a high level of statistical significance would perform at least as well as did the independent Bayesian model in the main experiment, with its crude data classification scheme.

Factors Influencing Length of Stay

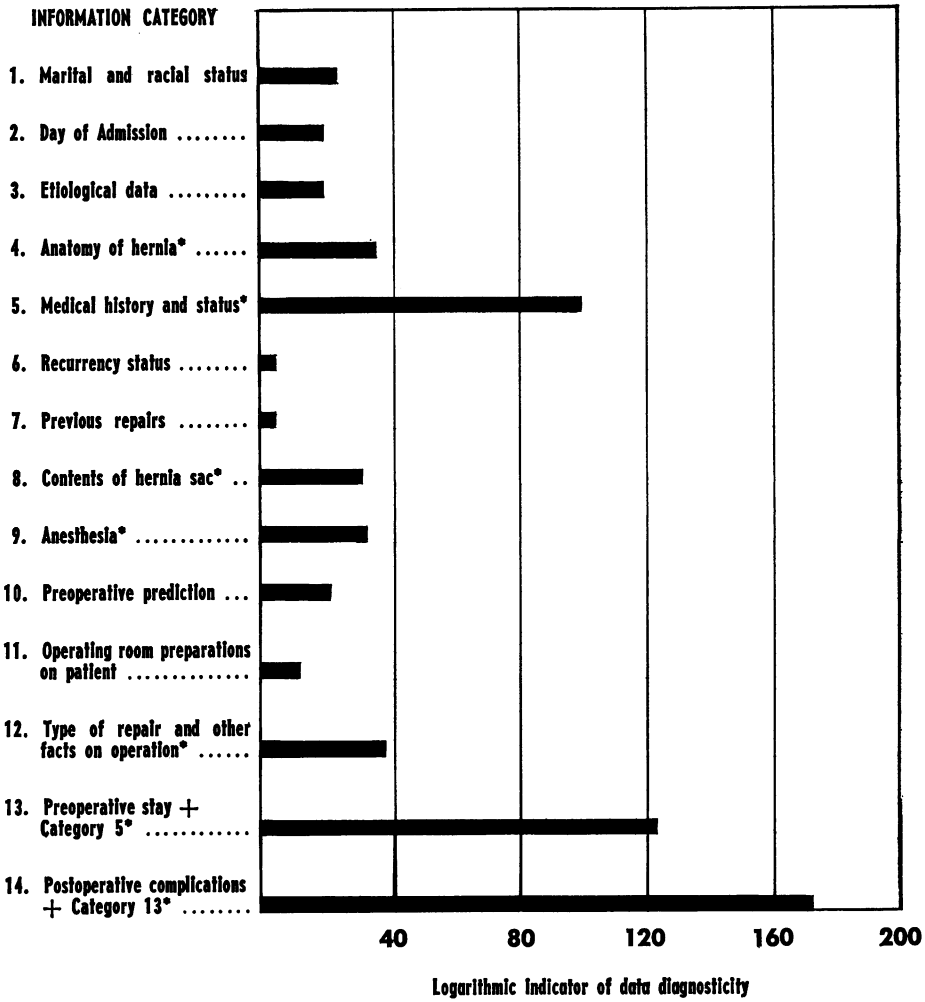
There have been numerous attempts to determine what factors influence length of stay and how they do so [24-28]. None have compared the descriptive results of empirical studies with the perceptions of physicians about the same questions. In this study, the regression analysis (RGN) and the subjective independent Bayesian (B.I.) methodologies were compared in this respect.

The degrees to which factors influenced the regression analysis equation were measured by the ratios of the factor coefficients to their standard errors. Table 6 lists those factors with ratios greater than 2:1. In the Bayesian approach the likelihood ratio is directly related to the impact that the data complex should have, from the physician estimator's point of view, on length of stay. Figure 6 presents data calculated by summing the natural logarithms of the 11 likelihood ratios associated with each data complex, as follows:

$$\sum_{i \neq B} \ln \frac{P(C_k \mid H_B)}{P(C_k \mid H_i)} \; .$$

Table 6. Factors that affect length-of-stay predictions in regression analysis (those factors for which the ratio of coefficient to standard error is greater than 2:1)

Preoperative Equation	*Operative Equation*
1—Age	1—Preoperative length of stay
2—Recurrency status	2—Recurrency status
3—ln (number of diagnostic procedures)	3—Number of drugs taken
4—Number of current medical problems	4—Day of admission
	5—Other operative procedures
Preoperative Equation (including demographic data)	*Postoperative Equation*
1—Day of admission	1—Preoperative length of stay
2—Recurrency status	2—Recurrency status
3—Number of preoperative diagnostic procedures	3—Day of admission
4—ln (number of preoperative consultations)	4—Type of work
5—Number of diagnostic procedures × number of preoperative consultations	5—ln (number of preoperative diagnostic procedures)
6—Type of compensation	6—Number of complications
7—Type of work	7—Type of hernia
8—Referral classification	
9—Type of hernia	



*Indicates that data have diagnostic value statistically different from zero at .99 level.

Fig. 6. Relative diagnosticity of data for likelihood ratio estimation (B.I.)

An analysis of the variances of these data showed that 7 of the 14 data complexes had a significant influence on length of stay, as indicated in the figure.

A review of the likelihood ratio estimates indicates that the physician estimators considered demographic characteristics as playing essentially no role in determining length of stay. The regression analysis, on the other hand, relied on several demographic factors in the preoperative equation; however, once more influential data became available, in the operative and postoperative equations, the demographic data played no significant role. One demographic factor, method of paying for medical care, has been of interest to several researchers, but showed very little effect in this study.

Because many factors were combined into one in the Bayesian approach and considered separately in the regression analysis, a thorough comparison of these methods cannot be made. Two areas of disagreement appear to exist, however. First, the physicians held that recurrency status did not play a major role in influencing length of stay; but the regression analysis indicated a strong relation. Second, the physicians held that no relation existed between length of stay and either of two factors: type of work done by the patient, or day of admission; but regression analysis indicated that both factors are important.

Discussion

This research indicates that subjective length-of-stay prediction is feasible under the conditions described here. Although all techniques were superior to the control historical mean (M), the Bayesian (B.I.) methodology appeared to perform best. It was superior to all other techniques by the M_c measure, superior to regression analysis by the D_m measure, and no worse than any other technique by both D_a and D_m measures.

Considering certain previous results, it is somewhat surprising that the advantage of the Bayesian methodology (B.I.) over the posterior odds technique (P.O.) was not more pronounced. One reason might be the lack of sensitivity in the D_a and D_m measures. The M_c measure and results of a later experiment [23] support this reason. The D_a and D_m measures implicitly assumed that it was good to minimize the deviation from the mean or mode; a more appropriate goal for the distribution estimators, of course, would be to describe accurately the true uncertainty inherent in a situation.

There are several reasons for the poor performance of the regression analysis. One reason is the small data base available to estimate the coefficients. There were not enough data to verify the coefficient estimates by using new data; additional data would have been helpful. However, later additional experiments that used over 1100 observations had similar results [23]. A second reason might be that the rather nondiagnostic variables entering the equation at the first two information levels did not explain enough variation to permit accurate prediction. Once a highly diagnostic variable (preoperative length of stay) enters, the performance of regression analysis greatly improves (Figure 3). It may be that subjective prediction techniques are more sensitive to data with low diagnostic value than is regression analysis.

The point estimates developed by the nonattending surgeon were based upon more medical experience and more thorough knowledge of the patient than were those estimates developed by the resident. That such knowledge did have some positive effects at higher information levels was suggested by Figure 4, but was not verified by the statistical analysis. It appears that neither the added experience of the surgeon nor the additional information available to the attending surgeon during the preoperative phase was sufficient to lead to a superior prediction. It could be that any additional information available to the attending surgeon was demographic rather than medical in nature; the analysis indicated that demographic information had low diagnostic value for length-of-stay prediction (Figures 2, 3, and 4).

The importance of conditional independence in using Bayes' Theorem was supported by the comparison of the independent and dependent Bayesian (B.I. and B.D.) methodologies (Figure 5; Tables 2 and 3). However, although conditional dependencies cannot be ignored, a comparison of the D_m results implies that detection of major dependencies may be sufficient.

Although the statistical analyses found no significant interaction between techniques and information levels (Table 2), there are some indications that such a tendency exists. Posterior odds (P.O.) and point estimates by residents (P.R.) appear to lose their superiority to the Bayesian model (B.I.) as the physicians are forced to aggregate additional information (Figure 3). This is consistent with the theory that the impact of the aggregation of large amounts of data creates conservatism.

Application to LOS Prediction

All the forecasting techniques appear to be more accurate predictors of length of stay than does the historical mean (M). However, potential applications hinge upon two additional questions. First, is it economically feasible to use any of the techniques? Second, will physicians participate in the estimation process on a continuing basis?

Cost can be measured in terms of physician time required to give the estimates. Times required for each technique are listed in Table 7. It is obvious that the attending surgeons and the regression equations can predict length of stay in the shortest time. It should be noted that the Bayesian time estimate is misleading; an operating system could greatly reduce physician estimation time by reusing the initial likelihood ratio estimates. In such case, physician participation would be required only when a new data complex

Table 7. Average physician time required to obtain a length-of-stay prediction

Forecasting Technique	*Time (Minutes)*	*Probability of a Correct Modal Estimate*
Regression analysis (RGN)	0	.093
Point estimate by resident (P.R.)	8	.177
Posterioi Odds (P.O.)	20	.239
Subjective, Conditionally Independent, Bayesian (B.I.)	30*	.375

*This value would be much smaller if the technique were used regularly.

occurred. However, even with these economies of scale, the average prediction time would probably not be as short as for attending surgeon estimates.

Because the Bayesian methodology has been shown to be the most accurate technique, and because the time it requires can be reduced, it has a place in length-of-stay estimation, especially when attending surgeon estimates cannot be used. The Bayesian model is a normative one, so it may have a strong role in studies requiring normative rather than descriptive statistics. Computer application to screening mechanisms for utilization review may be one example. The results reported here should be encouraging to those studying this area, because such a comparison between statistical and professional length-of-stay estimators is needed before designing a system to detect cases that are clinically deviant from normative practice.

Selection of a prediction methodology turns, in the final analysis, on the purpose for which the model is to be used and on the shape of the overall length-of-stay distribution. If the actual distribution has a small variance and is unimodal, a method that is accurate around the mode is desirable. When the distribution is flat, accuracy in the tails is desirable.

Many questions concerning length-of-stay predictions remain. The research indicates that the Bayesian model is the more accurate technique, but does not give it a resounding endorsement. One reason for this may be the simplicity of predicting length of stay for hernia operations. The herniotomy cases were chosen as a data source because of the wealth of information available. More conclusive results might be obtained by studying debilities characterized by a higher degree of uncertainty about length of stay.

Only eight patients were employed in this study. The experimental design was such that valid conclusions could be reached from a statistical point of view. However, a replication of this research using a larger and thus more representative sample would be valuable.

Training has a definite effect, to a degree not yet determined, on the performance of the estimators. Research on subjectively derived probabilities has shown that performance and knowledge of the data generating process are directly related [6]. Because the physicians were more familiar with the process of predicting length of stay when given certain data (Posterior Odds and Event estimation) than with that of predicting the data when given certain lengths of stay (Likelihood Ratio estimation), training may have been of more benefit in the latter task. Questions as to the amount and kind of training are of interest to all who apply subjective Bayesian decision processes. Questions concerning transfer of training are also of interest. Did the book bag and poker chip paradigm improve the estimator's performance?

Conditional independence continues to be a hindrance to the application of Bayesian statistics. How independent must the data be? How can conditional independence be detected? What role can investigators play in the detection process? What deflation or inflation of impact is caused by conditional dependence?

This research has demonstrated the applicability of subjective Bayesian models to medical decision problems. It indicates that hospital length of stay can be described, and that subjective models are applicable to real-world problems.

The Bayesian model has a potential in medical diagnosis, which has been investigated by several research teams [29-31]. When empirical data are lacking, this research indicates, subjective estimates may be substituted for the missing empirical data; further research into this question is in progress [32].

If doctors can describe their uncertainty about length of stay in numerical terms, it is probable that they could describe other phenomena in a similar fashion. Although nothing can replace the physician's judgment, Bayesian decision theory offers the physician another dimension in which to view such problems as deciding whether to operate in high-risk cases or to proceed with nonsurgical treatment.

References

1. Duncan, Acheson J. *Quality Control and Industrial Statistics.* Homewood, Ill. Irwin, 1962.
2. Bartscht, Karl G. Personal communication. Ann Arbor: The University of Michigan, June 1967.
3. Robinson, G. H. and H. Gargour. An analysis of hospital patient length of stay. Human Factors in Technology Research Group. Berkeley: University of California, March 1964.
4. Robinson, G. H., L. E. Davis and R. P. Leifer. Prediction of hospital length of stay. *Health Serv. Res. 1*:3, 287.
5. Young, J. *A queueing theory approach to the control of the hospital census.* Doctoral dissertation. Baltimore: Johns Hopkins University, 1962.
6. Phillips, L. D. *Some components of probabilistic inference.* Doctoral dissertation. Ann Arbor: The University of Michigan, 1965.
7. Kyburg, H. E., Jr. and H. E. Smokler (eds.), *Studies in Subjective Probability.* New York: John Wiley and Sons, Inc., 1964.
8. Savage, L. J. *The Foundations of Statistics.* New York: John Wiley and Sons, Inc., 1954.
9. Edwards, W. *Non-conservative probabilistic information processing systems.* ESD-TR-66-404, Institute of Science and Technology, Report 5893-22-F. Ann Arbor: University of Michigan, 1966.
10. Peterson, C. R. and A. J. Miller. Sensitivity of subjective probability revisions. *J. Exper. Psychol.*, 70, 11, 1965.
11. Phillips, L. D. and W. Edwards. *Conservatism in a simple probability inference task.* Report to Decision Sciences Laboratory, Electronic Systems Division Air Force Systems Command, AF19 (623)-2823, ESD-TR-65-217, April 1965.
12. Phillips, L. D., W. L. Hayes, and W. Edwards. Conservatism in complex probability inference. *IEEE Transactions on Human Factors in Electronics,* Vol. KFE-7, No. 1, 1966, p. 7.
13. Peterson, C. R. and A. J. Miller. Sensitivity of subjective probability revisions. *J. Exper. Psychol.*, 70, 11, 1965.
14. Schum, D. A., I. L. Goldstein, and J. F. Southard. Research on a simulated Bayesian information processing system. *IEEE Transactions on Human Factors in Electronics,* Vol. HFE-7, No. 1, 1966, p. 37.
15. Schum, D. A., et al. *Subjective probability revisions under several cost-payoff arrangements.* Human Performance Center. Columbus, Ohio State University, 1966.
16. Kaplan, R. J. and J. R. Newman. Studies in probabilistic information processing. *IEEE Transactions on Human Factors in Electronics,* Vol. HFE-7, No. 1, 1966, p. 49.
17. Bash, P. L. *The development and implementation of a medical information retrieval*

system. Master's thesis. Ann Arbor: The University of Michigan, 1964.
18. Draper, N. R. and H. Smith. *Applied Regression Analysis.* New York: John Wiley and Sons, Inc., 1967.
19. Gustafson, David H. *Comparison of methodologies for predicting and explaining hospital length of stay.* Doctoral dissertation. Ann Arbor: University of Michigan, 1966.
20. Mosteller, F. and D. L. Wallace. Inference and Disputed Authorship: *The Federalist.* Reading, Mass.: Addison-Wesley, 1964.
21. Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill, 1962, p. 374.
22. Scheffe, H. *The Analysis of Variance.* New York: John Wiley and Sons, Inc., 1961.
23. Gustafson, D. H. *Evaluation of probabilistic information processing in medical decision making.* To be published in *J. Org. Behavior and Human Perf.*
24. Balintfy, Joseph L. *Mathematical models and analysis of certain stochastic processes in general hospitals.* Doctoral dissertation. Baltimore: Johns Hopkins University, 1962.
25. Kolouch, Fred T. Computer shows how patient stays vary. *Modern Hospital 105*:5, 130, 1965.
26. Riedel, D. C. and T. B. Fitzpatrick. *Patterns of Patient Care.* Ann Arbor: University of Michigan Press, 1964.
27. Robinson, G. H., L. E. Davis and G. C. Johnson. The physician as an estimator of hospital stay. *Human Factors, 8*:3, 1966.
28. Abstracts of Hospital Management Studies, Vols. 1, 2, 3, 4. Ann Arbor: University of Michigan.
29. Toronto, A. F., L. G. Veasey and H. R. Warner, Evaluation of a computer program for diagnosis of congenital heart disease. *Prog. Cardiov. Dis.,* 1963, No. 5, p. 362.
30. Overall, J. E. and C. M. Williams. Conditional probability for diagnosis of thyroid functions. *J.A.M.A.* Feb. 2, 1963, No. 183, p. 307.
31. Nugent, C. A. The diagnosis of Cushing's syndrome in *The Diagnostic Process,* J. A. Jaquez (ed.) Ann Arbor: Malloy, 1964.
32. Gustafson, D. H., et al. *Subjective probabilities in medical diagnosis.* To be published.