Practical Machine Learning for Prognosis: A Case for AUC and Hepatitis Guideline Driven Inputs

Lihong Li[1], Danielle Baghernejad[1], Robert J. Gambrel[1], and Damian R. Mingle[1]

Intermedix, Nashville TN 37219, USA,
Lihong.Li@Intermedix.com, Danielle.Baghernejad@Intermedix.com,
Robert.Gambrel@Intermedix.com,Damian.Mingle@Intermedix.com
WWW home page: http://Intermedix.com

Abstract. In this paper machine learning was used to predict hepatitis-caused mortality in a way that is analytically robust and clinically relevant. Specifically, predictor variables in the dataset were used to generate new features based on clinical guidelines on test result thresholds and risk levels. The model was optimized to maximize area under the receiver operator curve, which arguably has more clinical relevance. Finally, models were tested for performance on a holdout sample that never served as part of any training set, serving as a proxy for expected out-of-sample performance. Using this guideline-inspired feature engineering and a model ensembling approach, we found that our particular model has an AUC of 98% and an accuracy of 96.77% in the holdout data. Based on this, our approach may serve as a first step in integrating machine learning models with clinical decision-making.

Keywords: machine learning, hepatitis, survival, prognosis, clinical guidelines

1 Introduction

Hepatitis is a costly disease, both in terms of lost quality of life and the cost of treatments. With recent advancements in treating chronic hepatitis, there is renewed interest in determining the point at which the cost of treatments are offset by improvements in quality-adjusted life years [22, 10, 31]. Additionally, calls for preemptive screening to allow for early detection and intervention have been made for both prison populations [21] and the general public [13].

Critical to these estimates is the ability to predict the survival of patients with hepatitis, both with and without treatment. To that end, many researchers have used the University of California at Irvine's [23] hepatitis dataset to train models that predict mortality remarkably well. As discussed more thoroughly in Section 2, models trained on this dataset frequently predict mortality with over 90% accuracy. A number of researchers have applied new, cutting-edge machine learning techniques to the data, and a small cottage industry has emerged in which

papers incrementally improve on prior results by using ever more sophisticated algorithms.

While model accuracy is an easily understood metric, we argue that the modeling approaches listed in Table 1 are wrong in using it as the primary benchmark for model success. Prediction accuracy may serve as a simple cross model comparison baseline, but it is insufficient if the desire is to use a model's prediction in a clinical scenario. For example, in the UCI hepatitis dataset, 123 of 155 patients live. A totally naive classifier could predict 'live' for every patient and have an accuracy of 79.4%. In practice, a medical professional needs to know a model's predicted classification, its confidence in that classification, and its ability to discriminate among the response classes. Model accuracy is inadequate in this sense - we contend that model Area under the Receiver Operating Curve (ROC) is a better measure of a model's predictive performance.

Furthermore, we incorporate features engineered based on clinical guidelines in our model. While previous approaches have engaged in a variety of feature engineering techniques, we have found minimal initial application of domain expertise in the generation of these features. It is possible that rules- and tree-based classification techniques would discover some of these rules and boundaries on their own. However, by including them before any algorithms touch the data, we ensure that clinically significant data is considered in all of our models. We also perform feature generation through association rules.

Finally, we note that we tune our model's hyper-parameters using ten-fold cross-validation on the training set but evaluate its performance on a 20% holdout sample. This practice is rare but gives much higher confidence about its future performance on new cases. Prior research has used both cross-validation results from the entire dataset and summaries of the model's success on the training data with no metrics to predict out-of-sample predictive power.

This paper proceeds as follows: In Section 2 we first survey the latest approaches in modeling the UCI hepatitis data and further highlight areas in which we believe the analysis should be extended. In Section 3 we then provide a brief clinical background on hepatitis's effects on the body and the significance of the test results provided in the UCI dataset. In Section 4 we discuss our modeling approach and measures of performance. The final section concludes and proposes further areas for research.

2 Previous Approaches and Performance Metrics Using UCI Hepatitis Data

The UCI Hepatitis dataset has been used many times in prior research, both as one example among many in demonstrating a new statistical model or algorithm [12, 27, 30] and as a dataset of particular interest for optimizing model performance due to its medical relevancy. A selection of hepatitis-focused machine learning papers is presented in Table 1. Among this group, approaches can be divided based

on whether they use all observations in the dataset or subset to only cases with complete data on included features. Though some results worked best on only a subset of data [33], more recent endeavors have yielded high predictive accuracy using the whole dataset [1,16,7,28,29,4,8,6,26].

Additionally, results differ in the method used to estimate a model's out-of-sample performance. Many papers use a standard 10-fold cross-validation technique. Others use a traditional 80-20 or 60-40 train-test split of the data. Of particular interest, we highlight the work of Chen et al [8] who use an 80-20 split and a 10-fold cross-validation to find the best hyper-parameters in the training set; they then deploy the model and evaluate its performance on the 20% holdout sample. As we explain more thoroughly in Section 4.2, this performance on holdout data is especially compelling and reassuring in terms of out-of-sample performance.

Finally, authors have various metrics available to demonstrate model performance. Across the papers listed in Table 1, model accuracy is the universal choice. While accuracy is a straightforward and easily understandable metric of a model's performance, it masks underlying uncertainty about edge cases as well as the model's confidence in its classification across the spectrum of predicted probabilities. From the perspective of a physician receiving a model's prediction of mortality on a new case, simply knowing that a model is correct 95% of the time in the aggregate is insufficient; she needs to know the confidence in a *specific case's* prediction, which would be better conveyed by knowing the model's classification ability along its ROC curve.

When optimizing model predictive accuracy, an analyst has two resources at his or her disposal. First, feature engineering can help create highly informative model inputs that have strong correlation with the output of interest. Second, the machine learning model itself can be as complex as desired, ranging from simple decision trees to highly sophisticated neural networks or evolutionary learning models. Both aspects of modeling are important and have their associated costs in time and computing power.

We argue that the feature-engineering aspect has been underappreciated in this context. Table 1 is rich with papers demonstrating cutting-edge algorithms that find relationships between the dataset's feature set and its death indicator. When features are engineered on top of those in the dataset, they are created through principal components analysis, dimensionality reduction, or data-driven association rules [29, 28, 8, 16]. Largely missing from them, however, is a discussion of how and why any input features were manipulated and expanded to produce their predictive results. By performing a variety of straightforward feature engineering transformations of the data, as well as coding cut-points and indicator variables based on medical guidelines (explained in Section 3) we are

able to achieve a predictive accuracy of 96.77% and an AUC of 0.98 using standard machine learning techniques.

Author	Accuracy	Performance Derivation	Data Coverage	Main Impact
Afif, Hedar, Hamid, and Mahdy[1]	98.75	10-fold CV	Full	Method (3SVM)
Polat and Gunes 2007[29]	94.12	10-fold CV	Full	Method (PCA-AIRS)
Polat and Gunes 2006[28]	92.59	10-fold CV	Full	Method (FS-AIRS)
Bascil and Temurtas[6]	91.87	10-fold CV	Full	Method (MLNN)
Chen, Liu, Yang, Liu, and Wang[8]	96.77	31 observation test sample	Full	Method (LFDA-SVM)
Dogantekn, Dogantekin, and Avci[16]	94.16	60 observation test sample	Full	Method (LDA-ANFIS)
Neshat, Sargolzaei, Toosi, and Masoumi[26]	93.25	Best performance of 500 tests on random 25% samples	Full	Method (CBR-PSO)
Calisir and Dogantekin[7]	96.12	Training dataset	Full	Feature Extraction (PCA), Method (LSSVM)
Sartakhti, Zangooei, and Moza-fari[33]	96.25	10-fold CV	80 Cases	Method (SVM-SA)

Table 1: All papers listed reported accuracy as the primary metric of model performance.

3 Medical Background

The liver is a complex organ that carries out the metabolism of carbohydrates, proteins, and fats. Hepatitis is inflammation of the liver, causing common signs and symptoms that include fatigue, flu-like symptoms, abdominal pain, loss of appetite, and yellow skin or eyes. Hepatitis B virus (HBV) and hepatitis C virus (HCV) are recognized as the most important sources of liver cirrhosis, the final stage of various chronic liver diseases [32].

Since some of the enzymes and end products of the metabolic pathway are sensitive to liver damage, they may be considered as biochemical markers of liver

dysfunction [20]. The liver enzymes, aspartate aminotransferase (AST, formerly called SGOT), alanine aminotransferase (ALT), gamma-glutamyl transferase, and alkaline phosphatase, are widely used to detect the liver damage. Bilirubin, albumin, and prothrombin time are the liver function test for evaluating patients with hepatitis. In general, an abnormal serum albumin or prothrombin time may be seen in patients with impaired hepatic synthetic function. Serum bilirubin measures the liver's ability to detoxify metabolites and transport organic anions to bile. Elevations of liver enzymes (SGOT and alkaline phosphatase) can reflect damage to the liver or biliary obstruction. In this hepatitis dataset, some of the biochemical markers, such as serum bilirubin, albumin, SGOT, alkaline phosphatase, and prothrombin time, are useful in predicting the mortality of patients with hepatic dysfunction.

Although the Child-Pugh score (or the Child-Turcotte-Pugh) was originally used to predict mortality during surgery, it is now commonly used to assess the prognosis of chronic liver disease. The Child-Pugh score evaluates five clinical measures of liver disease: bilirubin, albumin, prothrombin time, ascites, and encephalopathy [11].

Measure	1 point	2 points	3 points
Total bilirubin, mol/L (mg/dL)	<34 (<2)	34—50 (2—3)	>50 (>3)
Serum albumin, g/dL	>3.5	2.8—3.5	<2.8
Prothrombin time, prolongation (s)	<4.0	4.0—6.0	>6.0
Ascites	None	Mild	Moderate to severe
Hepatic encephalopathy	None	Grade I—II	Grade III—IV

Table 2: 5 clinical measures of the Child-Pugh score.

The Child-Pugh score has proved useful in estimating the prognostic index of survival for decades. Suarez et al.[34] conducted a study in 144 patients with liver cirrhosis and transplant candidates to evaluate Child-Pugh stages. The results showed that the Child-Pugh score and spontaneous bacterial peritonitis were independent predictors of survival. While not included in the dataset, these findings indicate that the Child-Pugh score may be valuable for the feature-engineering stage.

Beside liver enzymes and liver function tests, there are other variables associated with mortality caused by chronic hepatitis disease. For example, mortality is significantly related to gender. According to research by Weissberg et al. [37], women had less severe liver disease than men, based on the survival in chronic hepatitis B on 379 patients. Szpakowski and Tucker [35] also confirmed HBV-related mortality was four times more common in males than in females. Taylor et al. [36] systematically reviewed the literature and found 41 articles

suggesting that cirrhosis, higher HBV viral level, and male sex were consistently associated with a significantly increased risk of death and liver cancer.

The severity and poor prognosis of various liver diseases is also associated with age. Based on a study of 6,689 patients infected with HBV, Szpakowski and Tucker [35] also found that the mortality of HBV increased markedly with increasing age over 40 in males and over 50 in females. The death rate for these patients increased markedly with age, and approximately 40 percent of all deaths in subjects over the age of 40 were HBV-related. For the age-specific prognosis following spontaneous hepatitis B e antigen (HBeAg) seroconversion research in chronic hepatitis B, Chen et al. [9] compared prognosis on patient groups with different ages. They found that patients with HBeAg seroconversion before age 30 have an excellent prognosis.

The prognosis of liver damage, however, cannot be simply identified with reliability on the basis of a single assessment; serial testing is always helpful. Based on the survival research in chronic hepatitis B on 379 patients, the patients 40 or older with total bilirubin level of 1.5 mg/dl (25 mmol/L) or more, ascites, and spider nevi were identified at a higher risk of death [37]. De Jongh et al. [14] reported that the duration of survival was significantly associated with age, serum aspartate aminotransferase levels, presence of esophageal varices, and all five components of the Child-Pugh index. They also identified that age, ascites, and total serum bilirubin level were the most powerful prognostic indicators after multivariate analysis. To discover the interesting patterns among the variables ourselves, we applied association rules to identify strong rules/relations in the feature-engineering stage.

4 Modeling Approach

4.1 Metric Selection

Much of the previous research has emphasized model accuracy, the proportion of correct predictions, as the performance metric. We argue that accuracy alone does not convey a full picture of the model performance for several reasons. First, most classifiers produce a probability estimation along with a label of the class prediction and unfortunately, this information is not utilized when calculating accuracy. Accuracy does not consider the probability—only the label assigned—which leads to an easily overlooked subtle nuance: how does one assign a label?

If a label is assigned by whichever class has the higher probability, then the decision threshold to determine class labeling is 0.5. There are other approaches for selecting the appropriate decision value, such as picking a threshold to maximize the F1-score, precision, or recall, but all of them consist of relatively weighing whatever information the analyst deems important to consider. When separate experiments are run with accuracy as the metric, an unbiased comparison of model performance cannot be made unless the metric used to

determine the decision threshold is also reported and consistent among experiments.

Threshold aside, the metric can also be undermined when faced with imbalanced class sizes, and given the class sizes in the hepatitis dataset, this concern should not be overlooked. The signal in the majority class can overwhelm the minority class, and if this is not taken into account, a model favoring the majority class can appear to be a better classifier at first glance. Consider the following confusion matrices for a hypothetical dataset with 90 instances for one class and 10 for another. With such a difference in class sizes, even a naive classifier can appear to be a strong performer, as demonstrated in Table 3.

	Predicted Positive	Predicted Negative
Actually Positive	0	10
Actually Negative	0	90

Table 3: An example of an unintelligent majority class classifier.

While a useless classifier, it yields an accuracy of 0.9. A more intelligent classifier will likely have both false positives and false negatives, but accuracy as a single number is unable to convey this. Indeed, in Table 4, the model's accuracy is only 0.85 even though the model attempted to learn more of the underlying relationship. Thus, accuracy alone can be a misleading representation of the efficacy of a model when class sizes are not balanced.

	Predicted Positive	Predicted Negative
Actually Positive	5	5
Actually Negative	10	80

Table 4: An example of a learning classifier.

With these considerations in mind, we propose instead to measure a model on the AUC (Area under the Curve) of the ROC. Accuracy is based on a single decision threshold, whereas a ROC curve calculates all possible thresholds and graphs the false positive rate against the true positive rate. The AUC is the area under the resulting curve and can be interpreted as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example. The AUC can be calculated [24] as

$$AUC = \frac{S - \frac{n_0(n_0+1)}{2}}{n_0 n_1} \quad (1)$$

Where n_0 and n_1 are the numbers of positive and negative records respectively, and S is the sum of the ranks of all positive examples in the list.

As an example, the following table shows a ranked list with five positive examples and five negative examples. The labels are sorted in order of the probability, and we would expect a good classifier to have perfect separation between the class labels.

Actual Class Label	-	-	-	-	+	-	+	+	+	+
Rank of Label	1	2	3	4	5	6	7	8	9	10
Rank of Positive Label					5		7	8	9	10

Table 5: A table demonstrating expected AUC rank.

The AUC for these records is equal to

$$AUC = \frac{(5+7+8+9+10) - \frac{5*6}{2}}{5*5} = \frac{24}{25} \quad (2)$$

This calculation demonstrates that AUC is a measure for the quality of ranking; the more positive examples are ranked higher, the larger S becomes. Thus, it measures the classifiers skill in ranking a set of patterns according to the degree to which they belong to the positive class without actually assigning patterns to classes. From an actionability standpoint, we argue that this measure is more insightful: while decision makers may want to know how accurate the model is for a point estimate, the degree of confidence is also quite important. If a model has high accuracy but most labels were near the decision boundary and thus with a little variation could have been changed, decision makers may be less prone to accept a model over one that is more confident in the strength of its labeling.

4.2 Holdout Data for Generalization

In addition to AUC, we also employ a holdout set for additional confidence in true model performance. K-Fold cross validation is typically used to give an estimate of how well the model generalizes to new, unseen data. This is achieved by splitting the data into separate datasets, typically 5-10, calculating the error on each fold by using the other k-1 folds as training data and averaging across the folds for a final estimate on an accuracy metric.

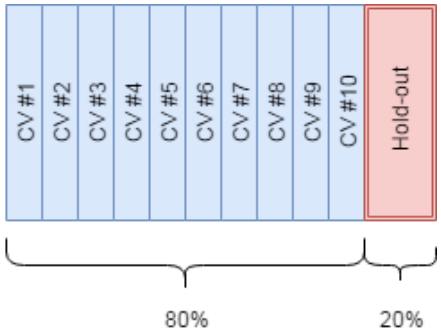


Fig.1: Percentage of data allocated to the hold-out set.

However, this is seldom done only once. Instead, this is repeated several times with different values of the model's hyper parameters. The final model is chosen by selecting the hyper parameters that perform best. While this does give an idea of the generalization error, since the hyper parameters are also tuned in this stage by, a true measure of generalization is rarely achieved.

While it is a subtle dependence, this can lead to over fitting to the holdout dataset. In practice, decision makers rarely want to learn from the past; they would rather deploy a model to predict on future events. For a truly useable model, then, model generalization must be prioritized. To that effect, we kept an additional holdout set to use for final reporting of model performance. We split the data into training, cross validation testing, and holdout testing, with the final metric calculated on the holdout dataset to represent how well we expect the data to generalize to future unseen observations.

4.3 Data set

The hepatitis dataset in this study is from UCI machine learning repository, which included 155 samples with 20 attributes (14 binary and 6 numeric attributes). The objective of this dataset is to identify or predict whether patients with hepatitis are alive or not (1 for deceased and 2 for alive). The binary attributes and numeric attributes of dataset have been shown in Tables 6 and 7.

As for the numeric attributes, bilirubin is the catabolic product of hemoglobin produced within the reticuloendothelial system and is released in an unconjugated form before being transported to the liver. In the liver, UDP-glucuronyltransferase conjugates the water-insoluble unconjugated bilirubin to glucuronic acid, and conjugated bilirubin is in turn excreted into the bile [17]. The higher levels of serum bilirubin are observed in viral hepatitis, hepatocellular damage or toxic liver injury [20]. Serum albumin is produced by hepatocytes in the liver and forms a large proportion of all plasma protein. Normal range of serum albumin in adults

is 3.5 to 5 g/dL. Hepatic synthesis of albumin is decreased in end-stage liver disease [19]. Therefore, serum albumin levels are useful in monitoring liver synthetic activity.

Aspartate aminotransferase (SGOT or AST) is one of the aminotransferases and is highly concentrated in the liver. Normal serum AST is 0 to 35 U/L. The liver cells spill the SGOT into the blood, increasing the enzyme levels in the blood when the liver is damaged (as in viral hepatitis). Therefore, SGOT is often used to help monitor potential liver damage from the hepatitis. A damaged liver, acute or chronic, eventually causes an increase in serum concentration of aminotransferases [19].

Alkaline phosphate (ALP) is an enzyme that transports metabolites across cell membranes. It presents in mucosal epithelia of the small intestine, proximal convoluted tubule of kidney, bone, liver, and placenta [20]. The most common causes of pathological elevation of ALP levels include liver and bone disease [18].

Binary Attribute	1	2	Missing
Class	Die: 32 (20.6%)	Live: 123 (79.4%)	0 (0.0%)
Sex	Male: 139 (89.7%)	Female: 16 (10.3%)	0 (0.0%)
Steroid	No: 76 (49.0%)	Yes: 78 (50.3%)	1 (0.7%)
Antivirals	No: 24 (15.5%)	Yes: 131 (84.5%)	0 (0.0%)
Fatigue	No: 100 (64.5%)	Yes: 54 (34.8%)	1 (0.7%)
Malaise	No: 61 (39.4%)	Yes: 93 (60.0%)	1 (0.7%)
Anorexia	No: 32 (20.6%)	Yes: 122 (78.7%)	1 (0.7%)
Liver Big	No: 25 (16.1%)	Yes: 120 (77.4%)	10 (6.5%)
Liver Firm	No: 60 (38.7%)	Yes: 84 (54.2%)	11 (7.1%)
Spleen Palpable	No: 30 (19.4%)	Yes: 120 (77.4%)	5 (3.2%)
Spiders	No: 51 (32.9%)	Yes: 99 (63.9%)	5 (3.2%)
Ascites	No: 20 (12.9%)	Yes: 130 (83.9%)	5 (3.2%)
Varices	No: 18 (11.6%)	Yes: 132 (85.2%)	5 (3.2%)
Histology	No: 85 (54.8%)	Yes: 70 (45.2%)	0 (0.0%)

Table 6: Details of binary attributes in hepatitis dataset

4.4 Feature Engineering

Much of the literature focuses on modeling based off the original data, focusing instead on fine tuning the algorithm. While important, feature engineering is a valuable process that allows for additional insight. Many contests on Kaggle, a machine learning competition website, credit creative feature engineering as a main component of the winning algorithms. Indeed, the highest scoring model

Numeric Attribute	Min	Max	Mean	Median	StdDev	Missing
Age	7	78	41.2	39	12.57	0 (0.0%)
Bilirubin	0.3	8	1.43	1.0	1.21	6 (3.9%)
Alk Phosphate	26	295	105.33	85	51.51	29 (18.7%)
SGOT (AST)	14	648	85.89	58	89.65	4 (2.6%)
Albumin	2.1	6.4	3.82	4	0.65	16 (10.3%)
Protime	0	100	61.85	61	22.88	67 (43.2%)

Table 7: Details of numeric attributes in hepatitis dataset

on the hepatitis dataset included over 200 variables engineered from rule mining to add additional nuance to the data. Feature engineering can range from basic transformations of the data to the creation of new variables utilizing additional information. For this study, both avenues were explored, with variable transformations being applied at the modeling stage, outlined in section 4.5.

For novel feature engineering, insight gained from medical literature was utilized. Since age seemed to be an important factor in prognosis, patients were grouped based on different age ranges with ≤30, 31-40, 41-60, and >60 in the dataset. New score variables of bilirubin, albumin, and ascites were also calculated based on Child-Pugh score. Given that the missing-ness rate of prothrombin time was high, approximately 43 percent, and hepatic encephalopathy was not included in the dataset, these variables were not included in the score calculations. Ascites was recorded as either 'yes' or 'no' in the data; we mapped 1 point for 'no' and 3 points for 'yes.' We also created new level variables of bilirubin, albumin, SGOT, and alkaline phosphate based on the value.

Binning of the numerical variables was also employed to group high and low levels together. Bilirubin was grouped with three levels based on the numerical ranges <1.2, 1.2-2.5, and >2.5. Albumin level was grouped into four categories, <3.0, 3.0-3.4, 3.5-4.5, and >4.5. The new SGOT-level variable was grouped at 0-40, 41-100, 101-199, 200-400, and >400. Finally, there were two levels for variable ALP-levels: <112 and 112-300.

To identify the potential interesting patterns among variables, we applied association rules in feature engineering. The Association Rules algorithm is a rules-based machine learning method to uncover how variables are associated to each other. The association rules are formally defined as [2]: Let $I = \{i_1, i_2, \cdots, i_n\}$ be a set of items, and $D = \{t_1, t_2, \cdots, t_m\}$ be a set of transactions, where each transaction t is a set of items such that $t \subseteq I$. A transaction t is said to contain X, a set of items in I, if $X \subseteq t$. An association rule is an implication of the form $X \rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \Phi$. It shows that the presence of X in a transaction will imply the presence of Y. Association rules are intended to identify strong rules in the dataset based on the measures of interestingness. Support, confidence, and lift are three

most common ways to measure association to select interesting rules from all possible rules.

Support is about how frequently an itemset is in the dataset, as measured by the proportion of transactions which contain *X*. It can measure the significance or importance of an itemset.

$$supp(X) = \frac{|\{t \in D; X \subseteq t\}|}{|D|} \tag{3}$$

Confidence is an indication of how likely item *Y* is purchased when item *X* is purchased. It is measured by the proportion of transactions with item *X* in which item *Y* also appears. Confidence gives different values for the rules $X \Rightarrow Y$ and $Y \Rightarrow X$.

$$conf(X \Rightarrow Y) = \frac{supp(X \Rightarrow Y)}{supp(X)} = \frac{supp(X \cup Y)}{supp(X)} \tag{4}$$

Support and confidence are evaluated to determine whether a rule should be kept in two steps. In the first step, support is used to find frequent (significant) itemsets that meet the minimum support constraint. Then confidence is used in the next step to produce rules from the frequent itemsets that exceed a minimum confidence constraint [2].

Lift measures how many times more often *X* and *Y* occur together than expected if they were statistically independent. If lift equals 1 for the rule, it implies that the probability of occurrence of the antecedent and that of the consequent are independent of each other, i.e., no association between items. A lift value greater than 1 means that item *Y* is likely to be bought if item *X* is bought and that those rules are potentially useful for predicting the consequent. A value less than 1 means that item *Y* is unlikely to be bought if item *X* is bought.

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)} \tag{5}$$

The algorithm produced thousands of association rules with support=0.02 and confidence=0.80 in this hepatitis dataset, which made it very difficult to analyze or identify those interesting/useful ones. To overcome this problem, another constraint of interest measures (chi-square) was also applied. Chi-square is used to test for independence between the left-hand-side (LHS) and right-hand-side (RHS) of the rule [25]. It is computed from a 2 by 2 table. For the rule $L \Rightarrow R$, the counts in this table are the number of transactions containing L and R; L but not R; not L but R; and not R and not L.

After pruning the discovered association rules with lift and chi-square threshold, we focused on the essential relationships and selectively created 24 relevant rules as our new variables. Part of the generated rules are listed in the table below.

Missing data was also taken into consideration, since much of the literature either imputes the missing values with the means or discarded rows or columns

with missing data all together. While a commonly used approach, this does not explore the potential significance of the data being missing in the first place. If there is a mechanism behind a variable being missing, such as a doctor preferring

lhs	rhs	support	confidence	lift	Chi-Square
Age group=41-60, Ascites=1	Class=1	0.065	0.909	4.403	4.374
Age group=41-60, Liver Firm=2, Spiders=1	Class=1	0.065	0.909	4.403	4.374
Malaise=1 ,Liver Big=2, Varices=1, Histology=2	Class=1	0.052	0.889	4.306	3.329
Ascites=1, Varices=1	Class=1	0.045	0.875	4.238	2.821
Age group=41-60, Sex=1, Varices=1	Class=1	0.058	0.818	3.963	3.315
Liver Big=2, Spiders=1, Ascites=1	Class=1	0.058	0.818	3.963	3.315
Liver Big=2, Spleen Palpable=2, Ascites=1, Histology=2	Class=1	0.058	0.818	3.963	3.315
Age group=41-60, Malaise=1 ,Anorexia=2, Histology=2	Class=1	0.058	0.818	3.963	3.315
Fatigue=1, Liver Big=2, Albumin level=<3.0,Histology=2	Class=1	0.052	0.800	3.875	2.818
Age group=41-60, Spiders=1, Alk Phosphate level=<112,Histology=2	Class=1	0.052	0.800	3.875	2.818

Table 8: Selected rules based on support, confidence, lift and Chi-square

not to run a test or not running a test if the results of another test deems it unnecessary, it can provide added insight as to why the variable was not recorded. Thus, we created binary indicator variables for each column with missing data and indicator variables for combinations of variables being missing. After calculating all combinations, columns containing redundant missing data patterns were removed, which resulted in 90 distinct missing data combinations.

4.5 Modeling

With the goal of finding both a highly accurate and highly usable model, a model featuring parsimony and robustness was chosen. A parsimonious model is defined as having coefficients with smaller absolute values as well as fewer non-zero coefficients, which is important for both issues of variable co-linearity and over-fitting. With a dataset this small, over-fitting is a concern since making a model too complex may lead to learning the data present instead of generalizing to future data.

In addition to building several separate, highly usable models, a final model consisting of an ensemble of three separate models was created. Since it is possible for one model to detect different nuances in a dataset over another, an ensemble of several uncorrelated models together can increase accuracy and reduce variance [3].

Elastic-Net Classifier The Elastic Net is an extension of logistic regression where the optimizer attempts to find a parsimonious model through regularization. While most classifiers emphasize either an L1 or and L2 penalty, Elastic Net is unique in that is utilizes both, with the degree of regularization for both being the two main hyper parameters that control the model. It favors fewer coefficients by implementing an L1 Lasso and favoring smaller coefficients by utilizing an L2 Ridge.

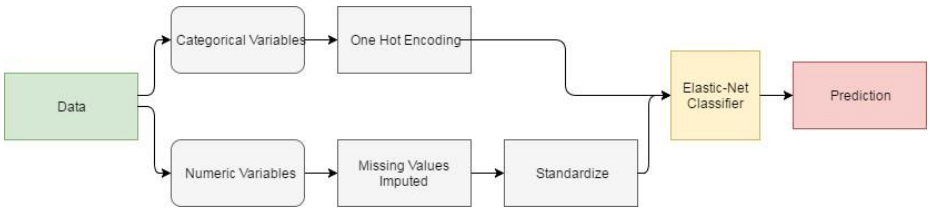


Fig.2: Machine learning work-flow for Elastic-Net classifier.

To implement, categorical variables were first transformed using one-hot encoding, which entails creating a binary indicator variable for each unique categorical value in the original variable. Numeric variables first had missing values imputed and then were standardized. Both the categorical and numeric outputs were sent to the Elastic-Net Classifier to generate predictions.

Nystroem Kernel SVM Support vector machines (SVM) are a class of maximum margin classifiers. They seek to maximize the separation they find between classes and can include a penalty function that allows misclassification of some observations for the sake of wider margins between the classes. As a result, this makes the support vector machine a very robust class of machine learning models. SVMs can also make use of a kernel function, which allows for a non-linear transformation of the data before fitting the SVM. These kernel functions can be very useful for transforming a non-linear problem into a linear domain.

Instead of implementing a kernel function directly, we make use of the Nystroem method, which is a general method for low-rank approximations of kernels. It achieves this by essentially sub-sampling the data on which the kernel is

evaluated. The advantage of using approximate explicit feature maps compared to the kernel trick is that explicit mappings can significantly reduce the cost of learning with large datasets. Kernel map approximations allow SVMs to run much faster and scale up to bigger datasets, but there is a small trade off in accuracy.

To implement, categorical variables were first transformed using one-hot encoding. For numeric variables, missing values were first imputed with the mean value and then two new variables were created for each, one representing the standardized values and the other representing the Ridit transformed values. The Ridit transform calculates the value's percentile and then normalizes the value such that the mean calculated for the reference population will always be 0 and the score will be between -1 and 1. Intuitively, the Ridit transform can be interpreted to be an adjusted percentile score.

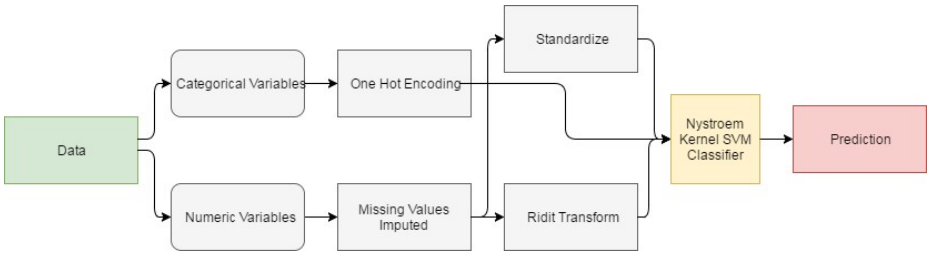


Fig.3: Machine learning work-flow for Nystroem Kernel SVM classifier.

Extra Trees Random forests are an ensemble method where hundreds of individual decision trees are fit to bootstrapped samples of the original dataset, with each tree being allowed to use a random selection of N variables, N being the major model hyper-parameter of this algorithm. Ensembling many re-sampled decision trees serves to reduce their variance, producing more stable estimators that generalize well out-of-sample. Random forests are extremely hard to over-fit, very accurate, generalize well, and require little tuning.

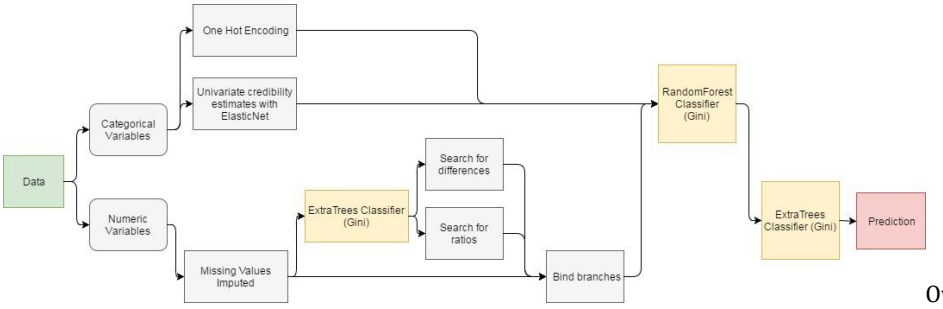
A further refinement of this method is the ExtraTrees model, which is a random forest with more randomness: the splits considered for each variable are also random. The ExtraTrees model has an additional advantage in that it is computationally very efficient: no sorting of the input data is required to find the splits because they are random.

To implement, additional feature engineering was employed to utilize the benefits of random forests. Since the algorithm runs quickly, this allows for identification of additional insightful variables by running the algorithm on variable subsets as an initial first pass. By running a first pass to filter the variables,

the feature space is reduced for efficient running of the final model. To perform, missing values of numerical variables were imputed with means, then ratios and differences of each pair of numeric values were calculated as new variables. These two sets of variables were then treated as inputs for an ExtraTrees classifier, which starts with a baseline model fit to the input datasets numeric variables. The model then loops through the generated variables, fitting a model to the numeric variables with the new feature added. If the variable improves the test set accuracy of the model, the variable is kept. The resulting new variables are then binded back to the original dataset for final modeling.

In addition, categorical values were transformed using one hot encoding. Credibility estimates of the categorical values were also calculated as additional variables. Credibility theory is a branch of actuarial science designed to quantify how unique a particular outcome will be compared to an outcome deemed as typical. For machine learning in particular, it can help reduce high-cardinality categorical features into more efficient representations, using one hot encoding on the transformed data.

The resulting variables were then supplied as input for a Random Forest model.



Overleaf

Fig.4: Machine learning work-flow for ExtraTrees classifier.

Ensemble Model For a final model, an ensemble of the individual models was generated. Ensembling is a useful tool for several reasons [15]. The first is statistical: an algorithm can be viewed as searching a space of hypotheses to identify the best hypothesis in the space. The statistical problem arises when the amount of training data available is too small compared to the size of the hypothesis space. Without sufficient data, the algorithm can find many hypotheses that give the same training accuracy. By ensembling all the accurate classifiers, the algorithm can average their votes and reduce the risk of choosing the wrong classifier.

The second reason for ensembling is computational: many algorithms work by performing some type of local search that may get stuck in local optima. In cases where there is enough training data so that the statistical problem is absent, it may

still be difficult computationally for the algorithm to find the best hypothesis. An ensemble constructed by running the local search from many different starting points may provide a better approximation to the true unknown function.

Thus, for our final model, an ensemble of the three individual classifiers was made. The probabilities of the classifiers was used as input variables to an Elastic Net algorithm with only an L2 penalty employed, and the resulting output served as a final prediction. While some of the metrics below varied, both the accuracy and AUC of the ensemble improved, thus demonstrating the benefit of an ensembling approach.

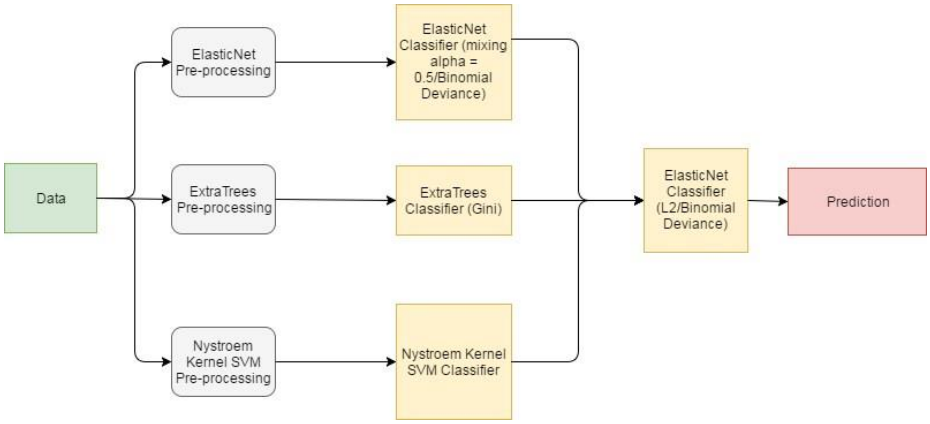


Fig.5: Machine learning work-flow for ensemble classifier.

	Elastic Net	Nystroem Kernel SVM	ExtraTrees	Ensemble
F1 Score	0.8333	0.8333	0.8	**0.9231**
True Positive Rate (Sensitivity)	0.8333	0.8333	0.6667	**1**
False Positive Rate (Fallout)	0.4	0.4	0	**0.04**
True Negative Rate (Specificity)	0.96	0.96	**1**	0.96
Positive Predictive Value (Precision)	0.8333	0.8333	**1**	0.8571
Negative Predictive Value	0.96	0.96	0.9259	**1**
Accuracy	0.9355	0.9355	0.9355	**0.9677**
Matthews Correlation Coefficient	**0.7933**	**0.7933**	0.7857	0.071

AUC	0.9567	0.9533	0.9467	**0.98**

Table 9: Results of approach broken out by individual contributing model and by blending the results in an ensemble.

5 Conclusion

We have expanded upon previous approaches to predict mortality in the UCI hepatitis dataset in a variety of ways. First, we used clinical guidelines to encode new features based on known thresholds for tests of enzymes and other biochemical markers in the blood. This feature generation ensures that medically important data are included in the models. By coding cutpoints and composite scores that are consistent with standard medical practice, we provide the models with an initial set of variables with the most predictive power possible.

Second, we used a variety of model families, an ensembling approach, and model evaluation using a holdout sample. This ensembling should reduce the risk of any single model overfitting the data and skewing predictions on new data. By testing our model's performance on a holdout dataset, we subject our model to a level of testing scrutiny not typically seen in prior work on this dataset. By following this rigorous method, we are much more confident in our model's ability to accurately classify future cases.

Finally, we have emphasized our model's performance on its AUC score, which we believe has more validity in measuring the model's power in an applied setting. In situations where the outcome's classes are unbalanced, as was the case in the hepatitis data, accuracy is an inadequate measure of model performance. AUC allows for more insight into the model's confidence of its classification of cases near the decision boundary. It is also less susceptible to manipulation of that decision boundary.

While we have focused our analysis on the prediction of hepatitis-related mortality using all the features in the dataset, we are aware that there is a crucial data element that we have excluded. The UCI dataset contains relative costs associated with each input feature; as our focus has been on making a model most useful for guiding medical practitioners, future work should include these variables. For example, the cost of labwork to obtain readings on bilirubin, AST, albumin, prothrombin time, and alkaline phosphate are substantially more than the costs to determine a patient's age, sex, and oral medical history. They also come with a time delay in delivery. However, they do share an initial fixed cost for a blood draw and processing, so there is a slight economy of scale in ordering multiple tests. Future work should examine whether all of these tests are necessary or whether a model's predictive power can be maintained without them. Moving forward, though, additional research into whether the missingness of these features is random will be needed.

Along a similar vein, additional provider feedback and integration will be necessary to make an analytic tool that is maximally effective in clinical practice. Any machine learning output will likely serve as one data consideration among many, subject to a provider's experience and expertise when determining a course of treatment. We have focused significant portions of our paper to discussing why models should be optimized for the measures that will be most meaningful to medical practitioners. In order to make the decision-making process more interactive, though, doctors should have input on how the model results are presented and also receive information about which features, currently with missing data on a given patient, would be most expected to improve predictive power if they are obtained. Such a system would transform a provider from a passive receiver of a machine's prediction to an active participant in determining which inputs the model receives; if a patient's prognosis is near the decision boundary and the doctor has the authority to order additional tests that may make the prediction more confident, the doctor may choose to delay an immediate treatment plan in favor of obtaining more data.

Finally, we observed that our dataset only contains patients with confirmed hepatitis. Given the cost of hepatitis treatment (and loss of quality of life if left undiagnosed), policymakers must consider whether general screening for hepatitis is warranted. Such concerns exist for the population overall but are of particular importance among the incarcerated population where hepatitis is both more prevalent per capita and also more successfully treated as patients are more likely to complete their full series of treatments [5]. By expanding the data to include patients with hepatitis across all stages of the disease, from initial diagnosis to liver failure, mortality can be predicted more successfully. In addition, variables indicating prior treatment plans and related illnesses that can indicate the severity of disease would make the models more powerful.

References

1. Mohammed H Afif, Abdel-Rahman Hedar, TH Abdel Hamid, and Yousef B Mahdy.Ss-svm (3svm): a new classification method for hepatitis disease diagnosis. *Int. J. Adv. Comput. Sci. Appl*, 4, 2013.
2. Rakesh Agrawal, Tomasz Imielin´ski, and Arun Swami. Mining association rulesbetween sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.
3. Kamal M. Ali and Michael J. Pazzani. Error reduction through learning multipledescriptions. *Machine Learning*, 24(3):173–202, Sep 1996.
4. Sana Ansari, Imran Shafi, Aiza Ansari, Jamil Ahmad, and Syed Ismail Shah. Diagnosis of liver disease induced by hepatitis virus using artificial neural networks. In *Multitopic Conference (INMIC), 2011 IEEE 14th International*, pages 8–12. IEEE, 2011.
5. EJ Aspinall, W Mitchell, J Schofield, A Cairns, S Lamond, P Bramley, SE Peters,H Valerio, J Tomnay, DJ Goldberg, et al. A matched comparison study of hepatitis c treatment outcomes in the prison and community setting, and an analysis of the impact of prison release or transfer during therapy. *Journal of viral hepatitis*, 23(12):1009–1016, 2016.

6. M Serdar Bascil and Feyzullah Temurtas. A study on hepatitis disease diagnosis using multilayer neural network with levenberg marquardt training algorithm. *Journal of medical systems*, 35(3):433–436, 2011.
7. Duygu C¸ali¸sir and Esin Dogantekin. A new intelligent hepatitis diagnosis system: Pca–lssvm. *Expert Systems with Applications*, 38(8):10705–10708, 2011.
8. Hui-Ling Chen, Da-You Liu, Bo Yang, Jie Liu, and Gang Wang. A new hybridmethod based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis. *Expert Systems with Applications*, 38(9):11796–11803, 2011.
9. Yi-Cheng Chen, Chia-Ming Chu, and Yun-Fan Liaw. Age-specific prognosis following spontaneous hepatitis b e antigen seroconversion in chronic hepatitis b. *Hepatology*, 51(2):435–444, 2010.
10. Jagpreet Chhatwal, Tianhua He, Chin Hur, and Maria A Lopez-Olivo. Directacting antiviral agents for patients with hepatitis c virus genotype 1 infection are cost-saving. *Clinical Gastroenterology and Hepatology*, 15(6):827–837, 2017.
11. E Cholongitas, GV Papatheodoridis, M Vangeli, N Terreni, D Patch, and AK Burroughs. Systematic review: the model for end-stage liver disease–should it replace child-pugh's classification for assessing prognosis in cirrhosis? *Alimentary pharmacology & therapeutics*, 22(11-12):1079–1089, 2005.
12. Suresh K Choubey, Jitender S Deogun, Vijay V Raghavan, and Hayri Sever. Acomparison of feature selection algorithms in the context of rough classifiers. In *Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on*, volume 2, pages 1122–1128. IEEE, 1996.
13. Stephanie Coward, Laura Leggett, Gilaad G Kaplan, and Fiona Clement. Costeffectiveness of screening for hepatitis c virus: a systematic review of economic evaluations. *BMJ open*, 6(9):e011821, 2016.
14. Felix E De Jongh, Harry LA Janssen, A Robert, Wim CJ Hop, Solko W Schalm,and Mark Van Blankenstein. Survival and prognostic indicators in hepatitis b surface antigen-positive cirrhosis of the liver. *Gastroenterology*, 103(5):1630–1635, 1992.
15. T Dietterich. Ensemble methods in machine learning. pages 1–15, 2000.
16. Esin Dogantekin, Akif Dogantekin, and Derya Avci. Automatic hepatitis diagnosissystem based on linear discriminant analysis and adaptive network based on fuzzy inference system. *Expert Systems with Applications*, 36(8):11282–11286, 2009.
17. Johan Fevery and Norbert Blanckaert. What can we learn from analysis of serumbilirubin? *Journal of hepatology*, 2(1):113–121, 1986.
18. William H Fishman. Alkaline phosphatase isozymes: recent progress. *Clinical biochemistry*, 23(2):99–104, 1990.
19. Edoardo G Giannini, Roberto Testa, and Vincenzo Savarino. Liver enzyme alteration: a guide for clinicians. *Canadian medical association journal*, 172(3):367–379, 2005.
20. Shivaraj Gowda, Prakash B Desai, Vinayak V Hull, Avinash A K Math, Sonal NVernekar, and Shruthi S Kulkarni. A review on laboratory liver function tests. *The Pan African Medical Journal*, 3, 2009.
21. Tianhua He, Kan Li, Mark S Roberts, Anne C Spaulding, Turgay Ayer, John JGrefenstette, and Jagpreet Chhatwal. Prevention of hepatitis c by screening and treatment in us prisonsprevention of hepatitis c in us prisons. *Annals of internal medicine*, 164(2):84–92, 2016.
22. Andrew J Leidner, Harrell W Chesson, Philip R Spradling, and Scott D Holmberg. Assessing the effect of potential reductions in non-hepatic mortality on the estimated

cost-effectiveness of hepatitis c treatment in early stages of liver disease. *Applied health economics and health policy*, 15(1):65–74, 2017.

23. M. Lichman. UCI machine learning repository, 2013.
24. Charles Ling, Jin Huang, and Harry Zhang. Auc: a better measure than accuracy incomparing learning algorithms. *Advances in Artificial Intelligence*, pages 991–991, 2003.
25. Bing Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discoveredassociations. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125–134. ACM, 1999.
26. Mehdi Neshat, Mehdi Sargolzaei, Adel Nadjaran Toosi, and Azra Masoumi. Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization. *ISRN Artificial Intelligence*, 2012, 2012.
27. David W Opitz and Richard Maclin. Popular ensemble methods: An empiricalstudy. *J. Artif. Intell. Res. (JAIR)*, 11:169, 1999.
28. Kemal Polat and Salih Gu¨ne¸s. Hepatitis disease diagnosis using a new hybrid system based on feature selection (fs) and artificial immune recognition system with fuzzy resource allocation. *Digital Signal Processing*, 16(6):889–901, 2006.
29. Kemal Polat and Salih Gu¨ne¸s. Prediction of hepatitis disease based on principalcomponent analysis and artificial immune recognition system. *Applied Mathematics and computation*, 189(2):1282–1291, 2007.
30. Michael L Raymer, Travis E Doom, Leslie A Kuhn, and William F. Punch. Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(5):802–813, 2003.
31. Homie Razavi, Antoine C ElKhoury, Elamin Elbasha, Chris Estes, Ken Pasini,Thierry Poynard, and Ritesh Kumar. Chronic hepatitis c virus (hcv) disease burden and cost in the united states. *Hepatology*, 57(6):2164–2170, 2013.
32. Marcia Samada and Julio C Herna´ndez. Prognostic factors for survival in patientswith liver cirrhosis. In *Liver Transplantation-Basic Issues*. InTech, 2012.
33. Javad Salimi Sartakhti, Mohammad Hossein Zangooei, and Kourosh Mozafari.Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (svm-sa). *Computer methods and programs in biomedicine*, 108(2):570–579, 2012.
34. M Samada Suarez, JC Herna´ndez Perera, L Ramos Robaina, L Barroso Ma´rquez,L Gonza´lez Rapado, M Cepero Vald´es, H Herna´ndez Rivero, A Abdo Cuza, A Roque Vald´es, J P´erez Bernal, et al. Factors that predict survival in patients with cirrhosis considered for liver transplantation. In *Transplantation proceedings*, volume 40, pages 2965–2967. Elsevier, 2008.
35. Jean-Luc Szpakowski and Lue-Yen Tucker. Causes of death in patients with hepatitis b: a natural history cohort study in the united states. *Hepatology*, 58(1):21–30, 2013.
36. Brent C Taylor, Jian-Min Yuan, Tatyana A Shamliyan, Aasma Shaukat, Robert LKane, and Timothy J Wilt. Clinical outcomes in adults with chronic hepatitis b in association with patient and viral characteristics: a systematic review of evidence. *Hepatology*, 49(S5), 2009.
37. Jed I Weissberg, Ljudevit L Andres, Coleman I Smith, Sheila Weick, Joanne ENichols, Gabriel Garcia, William S Robinson, Thomas C Merigan, and Peter B Gregory. Survival in chronic hepatitis b. *Ann Intern Med*, 101:613–616, 1984.