Predicting class membership using Imputation methods for clinical variable for hepatocellular carcinoma (HCC)

Amrina Ferdous[1], Nairanjana Dasgupta[2], Sayed Daoud[3]

1) Amrina Ferdous. (**Corresponding Author**)

Master's in Statistics, Department of Mathematics and Statistics,

Washington State University,

Pullman, WA 99164, USA.

Email: amrina.ferdous@wsu.edu

2) Nairanjana Dasgupta.

Professor, Department of Mathematics and Statistics, Washington State University,

Pullman, WA 99164, USA.

Email: dasgupta@wsu.edu

3) Sayed Daoud.

Center for Integrated Biotechnology, Washington State University,

Pullman, WA 99164, USA.

Email: daoud@wsu.edu

No. of Pages: 10 (excluding this page)

No. of Tables: 03

No. of Figures: 02

Running title: Predicting class membership using Imputation methods for clinical variable for hepatocellular carcinoma (HCC)

Predicting class membership using Imputation methods for clinical variable for hepatocellular carcinoma (HCC)

Amrina Ferdous[1], Nairanjana Dasgupta[2], Sayed Daoud[3]

Abstract: It is an accepted fact that infection with Hepatitis C virus (HCV) is a leading risk factor for chronic liver disease progression, including cirrhosis and hepatocellular carcinoma (HCC). Prevalence rates of HCV infection range from 2% in the US to as high as 14% in developing countries such as Egypt. In an attempt to define the molecular signatures of HCV-induced HCC, the methylation signatures in Egyptian tissue samples from patients with active HCC, patients with HCV and normal liver tissues were compared using a panel of genes that are commonly hyper methylated in other solid tumors. The prognostic impact of the aberrant promoter methylation (PM) status of genes was also correlated to the clinicopathological parameters of patients. However, in most cases, clinico-pathological variables are not available for normal patients. This kind of situation is common in diagnostic study of cancer, as often the "normal" group is constructed from donor data; hence, clinical information is not readily available. What we did was solve this problem our two step algorithm for data imputation. Using our imputation methods the probability of correct class prediction increased from 60% to 72%. Similarly the R-square increased from 69% to 79%. We believe this approach can help other scientists as who are facing the problem of missing data as well. While the idea of imputation is not new, we think our two step approach is novel and potentially applicable in other cancer studies as well.

Keywords: Linear regression, logistic regression, R-square, p-value, percent concordant, Akaike Information Criterion, Bayesian Information Criterion, Imputation, Hepatocellular carcinoma (liver cancer).

---

1) Amrina Ferdous.
Master's in Statistics, Department of Mathematics and Statistics, Washington State University,
Pullman, WA 99164, USA.
Email: amrina.ferdous@wsu.edu

2) Nairanjana Dasgupta.
Professor, Department of Mathematics and Statistics, Washington State University,
Pullman, WA 99164, USA.
Email: dasgupta@wsu.edu

3) Sayed Daoud.
Center for Integrated Biotechnology, Washington State University,
Pullman, WA 99164, USA.
Email: daoud@wsu.edu

1. Introduction:

The liver continuously filters blood that circulates through the body, converting nutrients and drugs absorbed from the digestive tract into ready-to-use chemicals. The liver performs many other important functions, such as removing toxins and other chemical waste products from the blood and readying them for excretion. Because all the blood in the body must pass through it, the liver is unusually accessible to cancer cells traveling in the bloodstream. The liver can be affected by primary liver cancer, which arises in the liver, or by cancer which forms in other parts of the body and then spreads to the liver.

Hepatocellular carcinoma (HCC, also called malignant hepatoma) is the most common type of liver cancer worldwide and one of the leading cause of cancer-related death [1]. It accounts for approximately 600,000 deaths per year [2]. Most cases of HCC are secondary to either a viral hepatitis infection (hepatitis B or C) or cirrhosis (alcoholism being the most common cause of liver cirrhosis).

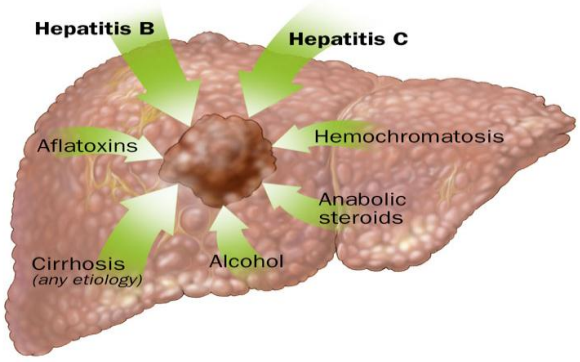

Figure-1: Affected human liver with hepatocellular carcinoma.

HCC is a relatively uncommon cancer in North America and Europe. In countries where hepatitis is not common, most cancers of the liver are not primary HCC but metastasis (cancers spread from elsewhere in the body such as the colon). The risk factors which are most important; varies widely from country to country. In Africa and Asia Hepatitis B will be the predominant cause of Hepatocellular Carcinoma. Whereas in countries, such as the United States, where Hepatitis B is rare because of high vaccination rates, the major cause of HCC is Cirrhosis (often due to alcohol abuse). In Egypt the incidence of HCC has doubled in the past 10 years, consequently it is now the second most incident and lethal cancer in men after lung cancer [1].

This study is a part of a larger on-going study looking at gene methylation for the different stages of liver disease. In a past study [9] the writers showed that the progression of liver disease from normal tissue to cancer went through various stages. The stages are no-disease, hepatitis B or C positive, cirrhosis and finally HCC. To understand this progression we have tissue samples on four stages Asymptomatic (people

with advanced liver disease but have had a liver transplant, we call this A), Chronic Hepatitis B and C (people living with Hep B or C, we call this group C), people with active HCC and liver tumor (call this group T) and people without disease, our normal group (called B, tissue collected from donors). From the tissue we have genetic information. However, the clinical information is only available for the groups A,C and T. The goal of the larger study was to see which variables (genetic and clinical) were predictors of these stages. However, we were hampered by not having clinical variables for the B group. The goal of this prospective study is trying to impute clinical information for group membership. The bigger aim of this study is to contribute to diagnostic study of cancer by imputing (generally unavailable) clinico-pathological variables for data taken from donors.

2. Methods:

2.1 Data: The data was collected from virology and immunology unit, cancer biology department, national cancer institute, Cairo University, Egypt.

Mainly two types of variables were present in the data set:

1) **Information from Gene methylation**: We have information on methylation or not from the following genes APC, P15, P73, P14, P16, DAPK, RAR, RASSF1A, O6MGMT, E_cadherin. This data is binary in nature. This data is available for the entire data set.
2) **Information from the Clinical variables**: We have several clinical variables for liver related variables. This information is available only on the patients from whom live tissue was extracted.

Tissue samples from 31 patients with HCC (T group), 38 from chronic hepatitis (HCV) patients. These 38 tissue samples were divided into two groups: 18 samples from patients with liver transplant (A group) and 20 samples that didn't (C group). For these three groups (A, C and T) we have the data for the eleven gene activation status as well as their clinicopathological information [9]. The data for the normal hepatic tissue was collected from liver donor samples and as such no clinical information is available for this group (called group B).

2.2 Description of Different stages of HCC: As stated above we have four stages of disease 'A' for asymptomatic, 'B' for normal (normal liver), 'C' for chronic Hepatitis B & C, 'T' for tumor (cirrhosis).

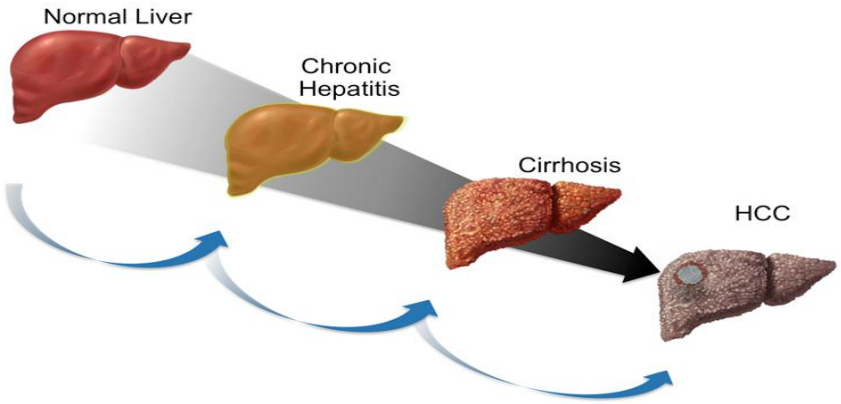

Figure-2: Different stages of Hepatocellular carcinoma (HCC) in human body.

Normal liver: A multilobed highly vascular reddish-brown glandular organ occupying most of the upper right part of the human abdominal cavity immediately below the diaphragm. It secretes bile, stores glycogen, detoxifies certain poisons, and plays an important part in the metabolism of carbohydrates, proteins, and fat, helping to maintain a correct balance of nutrients.

Chronic hepatitis B & C: The most common risk factor for liver cancer is chronic infection with the hepatitis B virus (HBV). Individuals chronically infected with HBV are 100 times more likely to develop liver cancer than uninfected people because the virus directly and repeatedly attacks the liver, which over time can lead to progressive liver damage and liver cancer. Hepatitis can have non-infectious causes too, including heavy drinking, drugs, allergic reactions, or obesity.

Cirrhosis: Cirrhosis is a slowly progressing disease in which healthy liver tissue is replaced with scar tissue, eventually preventing the liver from functioning properly. The scar tissue blocks the flow of blood through the liver and slows the processing of nutrients, hormones, drugs, and naturally produced toxins. It also slows the production of proteins and other substances made by the liver. According to the National Institutes of Health, cirrhosis is the 12th leading cause of death by disease.

Hepatocellular Carcinoma (HCC): The most common type of liver cancer, hepatocellular carcinoma (HCC), is a primary malignancy of the liver and occurs predominantly in patients with underlying chronic liver disease and cirrhosis.

Given the type of data we have, our class prediction and feature selection can be done in two different ways: using just the genetic information that we have for all four groups (A, C, T and B) or both the genetic information and clinical information, which is possible for the groups A, C and T. It is worthy to note that the missing clinical information isn't just for a few observations but for the whole B group. This type of missing information, where information for a whole group is missing, is potentially a common problem when donor tissues are used for the "normal" group. In retrospect, it makes sense, as liver biopsy information would not be available for anyone without significant hepatic issues. It would be non-ethical and extremely invasive to have this done for a person without hepatic disease issues. Hence, there is a need for understanding how to impute data in this case scenario.

3. Analysis:

Our methodology is summarized by the following 5-step algorithm:

Step 1: To compute nominal logistic regression with all (four) groups and only predictors with complete information to identify prediction statistics, correct predictions max-rescaled R-square.

Step 2: To compute step-wise logistic regressions for the categories with complete information (in our case A, C, T) as the nominal response and the gene activation status and clinical information as predictor variables. Select our list of variables using an alpha cut-off value of 0.15. This step would produce a list of predictors for both gene activation status and clinicopathological variables.

Step 3: To use the clinicopathological variables selected in Step 2 and the gene activation variables, to create a predictive model for the missing clinicopathogical variables for the entire data set.

Step 4: To compute a nominal logistic regression with all 4 categories using the predicted variables in Step 4 for the missing relevant (from step 2) clinicopathogical variables.

Step 5: To determine if there is increase in correct prediction and maximum re-scaled R-square over the model without the imputed values to ensure imputation is relevant in this case.

4. Results:

Step 1: Predicting all 4 (A, B, C, T) classes with ONLY gene activation status.
The model selects *RASSF1A*, *p73*, *p14*, *p16*, *OGMGMT* and *APC* as the pertinent features. The R-square for this model is 69% and the concordance is 88%. The observed versus predicted values is shown in Table 1.

Table 1: Observed and Predicted classes for A, B, C and T without clinical variables

	PREDICTED					TOTAL
OBSERVED		A	B	C	T	
	A	12	0	5	1	18
	B	12	0	1	0	13
	C	2	2	9	7	20
	T	0	0	3	28	31

Findings: The data for this model show a various limitation. It has low Max-rescaled R-square. Almost all the observations in the B group (normal) is misclassified into A group (asymptomatic hepatitis) with 1 in the C group (chronic hepatitis).

Step 2: Finding the relevant clinicopathogical variables using Step-wise Logit for A, C and T:

The selected model has the features: *RASSF1A*, *p73*, *E-cadherin* and platelets. In this case, platelets count is the only variable selected from the clinicopathological group and is the most significant (first to enter the step-wise regression list and significant at <.0001 level). While not strictly relevant, the R-square is 79.5% and the concordance is 96%. In this step we conclude that:

- The clinical variable platelets is the only relevant predictor for this model.
- The use of platelets improves our model significantly.

Step 3: Model Selection for a selected clinicopathological variable, platelets:

Using the linear model for platelets we predicted platelets, based on gene activation predictors and thus could impute the values of the platelets for all four groups.

Step 4: Nominal Logit with imputed platelets for Group B.

We used the imputed platelets and performed a feature selection for the 4 groups, including the gene information and the "imputed" platelets for the B group. The selected features, again include platelets, *p14*, *p16*, *Rassfa1a* and p73. This model had R-square of 79% and a concordance of 89.4%, as shown in Table 2.

Table 2: Observed and Predicted classes for A, B, C and T with imputed clinical variable

	PREDICTED					TOTAL
OBSERVED		A	B	C	T	
	A	6	3	5	0	14
	B	5	7	1	0	13
	C	3	2	12	1	18
	T	0	0	1	26	27

Step 5: Comparison of the models in Step 1 (without imputed clinicopathological variables) and Step 4 (with imputed Platelets)

Table 3: Side by side comparison of models in Step 1 and Step 4

	Without Imputed values (Step 1)	With Imputed values (Step 4)
R-square	69%	79%
Concordance	88%	90%

% of correct predictions for A, B, C and T	67%, 0%, 45%, 90%	45%, 54%, 67%, 97%

As shown in Table 3, the imputation did improve the model appreciably. For B, C and T the percentage of correct predictions has improved. This indicates that imputation was indeed a good option for this classification model.

5. Discussion:

In Table 3, we showed that using imputed values of clinical variables when such information is missing substantially enhanced our model. While R-square (here we used the max-rescaled R-square), AIC, BIC and concordance are all measures of the goodness of fit, what we are really interested is the class prediction, and to ensure that we do not have consequential false positives (predicting tumor, T, for a person from the normal, B group), or false negatives (predicting normal, B, when the person has a tumor). Hence, to understand this approach, we need to have a closer look at Tables 1 and 2. In Table 1, none of individuals who were initially considered normal were correctly classified into the normal class. Out of the 13, 12 were classified in the asymptomatic hepatitis (A group) and 1 in the chronic hepatitis, C, group. When we used the imputed platelets this results changed in Table 2: 7 are classified correctly and 5 are misclassified to A group and 1 in C group. This is a substantial improvement over the results in Table 1. Similarly, in Table 1, for the Tumor (T group), out of 31, 3 are misclassified to the chronic hepatitis C group, whereas in Table 2 out of 27, only 1 is misclassified into that group. In Table 1, 49 out of 82 (60%) is correctly predicted and in Table 2, 51 out of 72 (71%) is correctly predicted. This shows that imputing the clinical variables did improve the class prediction substantially.

In summary, in this report we encountered a situation where the original data for normal patients had no clinicopathological variables, while we had it for the other patients who undergone liver biopsy. This kind of clinical situation is common in diagnostic study of cancer, as often the “normal” group is constructed from donor data; hence, clinical information is not readily available. In this case we were able to solve the lack of clinicopathological data using a 5-step algorithm. We believe this approach can

help other scientists/clinicians who may encounter a similar problem of missing data as well. While the idea of imputation is not new or original, we believe that our two step imputation approach is novel and potentially applicable in other cancer studies as well. We can also conclude that an analysis considering predicted values for platelets would be a good idea rather than excluding this variable just because we have missing observations for normal people group.

References:

[1] Abdel-Rahman N. Zekri et. al. Methylation of multiple genes in hepatitis C virus associated hepatocellular carcinoma, Journal of Advanced Research, Cairo University, Egypt 2012.

[2] Lehman EM, Wilson ML. Epidemiology of hepatitis viruses among hepatocellular carcinoma cases and healthy people in Egypt: a systematic review and meta-analysis. Int J Cancer 2009; 124(3):690-7.

[3] Yeh MM, Yeung RS, Apisarnthanarax S, Bhattacharya R, Cuevas C, Harris WP, Hon TLK, Padia SA, Park JO, Riggle KM, Daoud SS**.** Multidisciplinary perspective of hepatocellular carcinoma: A Pacific Northwest experience. World J Hepatol 7:1460-83, 2015

[4] Zekri NA, Bahnasy AA, Shoeab FEM, Mohamed WS, El-Dahsan DH, Ali FT, Sabry GM, Dasgupta N, Daoud SS. Methylation of multiple genes in hepatitis C virus-associated hepatocellular carcinoma. J Advanced Res 5:27-40, 2014

[5] Boccaccio V and Bruno S. Management of HCV patients with cirrhosis with direct acting antivirals. Liver Int. 34:38-45, 2014

[6] Reau NS and Jensen DM. Sticker shock and the price of new therapies for hepatitis C: Is it worth it? Hepatology 59:1246–1249, 2014

[7] Michael E. DeBakey. Epidemiology of hepatocellular carcinoma in USA. Hepatol Res, September 2007; Suppl 2:S88-94.

[8]T F Greten, et. al. Survival rate in patients with hepatocellular carcinoma: a retrospective analysis of 389 patients (2009).

[9] American Cancer Society, Cancer Facts & Figures 2016. Atlanta, Ga: American Cancer Society; 2016.

[10] Chou R, Hartung D, Rahman B, Wasson N, Cottrell EB, Fu R. Comparative effectiveness of antiviral treatment for hepatitis C virus C infection in adults: a systematic review. Ann Intern Med. 158:114-23, 2013.