International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 36 - 41, 2015, ISSN:- 2319–8656

Design of a Clinical Decision Support System Framework for the Diagnosis and Prediction of Hepatitis B

Adekunle Y.A
Department of Computer Science,
Babcock University,
Ilishan-Remo, Ogun State,
Nigeria

Abstract: This paper proposes an adaptive framework for a Knowledge Based Intelligent Clinical Decision Support System for the prediction of hepatitis B which is one of the most deadly viral infections that has a monumental effect on the health of people afflicted with it and has for long remained a perennial health problem affecting a significant number of people the world over. In the framework the patient information is fed into the system; the Knowledge base stores all the information to be used by the Clinical Decision Support System and the classification/prediction algorithm chosen after a thorough evaluation of relevant classification algorithms for this work is the C4.5 Decision Tree Algorithm with its percentage of correctly classified instances given as 61.0734%; it searches the Knowledge base recursively and matches the patient information with the pertinent rules that suit each case and thereafter gives the most precise prediction as to whether the patient is prone to hepatitis B or not. This approach to the prediction of hepatitis B provides a very potent solution to the problem of determining if a person has the likelihood of developing this dreaded illness or is almost not susceptible to the ailment.

Keywords: Hepatitis , Clinical Decision Support System (CDSS), Medical Decision Support System (MDSS), Artificial Intelligence (AI), K Nearest Neighbor (K-NN), Decision Trees (DT), Support Vector Machine (SVM) and Sequential Minimal Optimization (SMO)

1. INTRODUCTION

In recent times, the development of intelligent decision making applications is fast gaining ground. This concept is known as Artificial Intelligence (AI). Artificial Intelligence has different sub-fields which include expert systems, machine vision, machine learning and natural language processing amongst others.

A Decision Support System is an interactive computer-based system intended to help decision makers utilize data and models in order to identify and solve problems and make decisions [1]. According to the Clinical Decision Support (CDS) Roadmap project, CDS is "providing clinicians, patients, or individuals with knowledge and person-specific or population information, intelligently filtered or present at appropriate times, to foster better health processes, better individual patient care, and better population health."

A Clinical Decision Support System (CDSS) is an active knowledge system, where two or more items of patient data are used to generate case-specific recommendation(s) [2]. This implies that a CDSS is a decision support system (DSS) that uses knowledge management to achieve clinical advice for patient care based on some number of items of patient data. This helps to ease the job of healthcare practitioners, especially in areas where the number of patients is overwhelming.

Hepatitis B is a viral disease process caused by the hepatitis B virus (HBV). The virus is endemic throughout the world. It is shed in all body fluids by individuals with acute or chronic infection. When transmission occurs from mother to child or between small children during play, the infection nearly always becomes chronic. By contrast, when transmission occurs in adolescents/adults—usually via sexual contact, contaminated needles or other sharp objects, and less often from transfusion of blood products—the infection usually resolves unless the individual is immunocompromised.

Two billion people worldwide have serologic evidence of past or present HBV infection, and 350 million are chronically infected and at risk of developing HBV-related liver disease. Some 15–40% of chronically infected patients will develop cirrhosis, progressing to liver failure and/or HCC. HBV infection accounts for 500,000–1,200,000 deaths each year.

Health-care workers remain an at-risk group due to the risk of needlestick injury, and they should therefore all be vaccinated before employment.

Individuals chronically infected with HBV are at increased risk of developing cirrhosis, leading to hepatic decompensation and hepatocellular carcinoma (HCC). Although most patients with chronic HBV infection do not develop hepatic complications, there is the likelihood that serious illness can develop during their lifetime, and it is more likely to occur in men.

Every individual chronically infected with HBV provides an avenue for further cases to be prevented. It is expedient to take the time needed to educate patients and to explain the risks that the infection poses to the patients themselves and to others.

Hepatitis B vaccination is highly efficacious, and universal vaccination at a tender age is desirable. At the very least, vaccination should be offered to all individuals who are at risk. Pregnant women ought to be screened for hepatitis B before delivery, as this offers an opportunity to prevent another generation of chronically infected persons.

HBV-related liver injury is majorly caused by immune-mediated mechanisms, mediated via cytotoxic T-lymphocyte

lysis of infected hepatocytes. The precise pathogenic mechanisms responsible for the HBV-associated acute and chronic necroinflammatory liver disease and the viral and/or host determinants of disease severity have only recently been established [3]. The immune response of the host to HBV-related antigens is important in determining the result of acute HBV infection. The strength of the immune response of the host is critical for clearing the virus, but this simultaneously causes liver injury (i.e., a form of "hepatitis" manifested by a rise in transaminases occurs before clearance of the virus can be achieved). Those who become chronically infected are not able to sustain an immune response to HBV and thus undergo intermittent episodes of hepatocyte destruction (hepatitis).

Most studies of acute HBV infection are only initiated after the onset of symptoms, so that the critical early events following HBV infection go without notice. A recent study serially profiled the genomic changes during viral entry, spread, and clearance of the virus and showed that HBV does not induce any interferon-regulated genes in the early phase of the infection. Additionally, no genes were up-regulated or down-regulated in the lag phase of the infection or during the phase of viral spread. This suggests that HBV may not induce the intrahepatic innate immune response. Thus, HBV may be a "stealth" virus early in the infection.

When neonates are infected during childbirth if their mother is HBeAg-positive, immune tolerance is induced as the fetus becomes tolerized to the e antigen, a soluble viral protein that crosses the placenta in utero. This immune-tolerant phase goes on for years to decades. Children born to mothers who are HBeAg-negative but have ongoing viral replication more often develop an acute hepatitis in the neonatal period, which is cleared by the infant. However, the infectivity of many women who are HBeAg-negative is often very low, so that only about 20% transmit hepatitis B to their offspring.

Table 1. Acute hepatitis B infection: the risk of chronicity is related to age at primary infection (Source: Neth, 2006)

Outcome	Neonates	Children	Adults
Chronic infection	90%	30%	1%
Recovery	10%	70%	99%

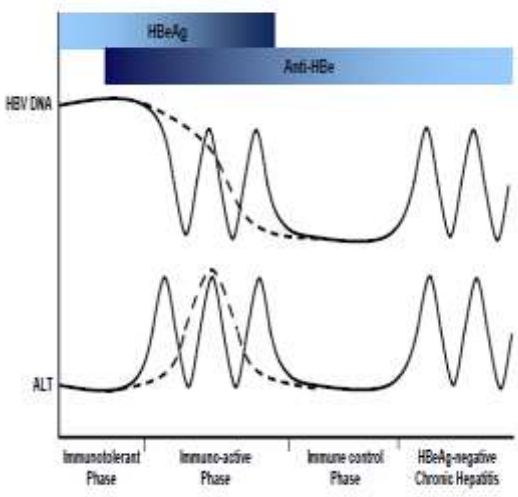

Figure 1. Chronic hepatitis B infection: phases of infection (Source: Buster, 2006)

Most cases of chronic hepatitis B in the reactivation phase are HBeAg-negative, but a few patients may be HBeAg-positive (Figure 2). The rates of progression to cirrhosis and hepatocellular carcinoma, with the associated mortality rates, are shown in Figure 2.

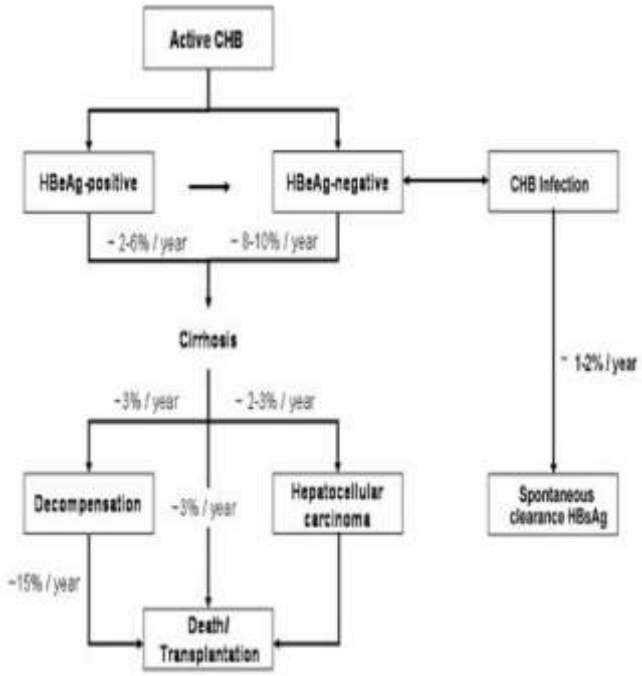

Figure 2. Progression to cirrhosis and hepatocellular carcinoma, with mortality rates (Source: Hepatol et al, 2003)

2. RELATED WORKS

2.1 Decision Support System for Heart Disease Based on sequential Minimal Optimization in Support Vector Machine

Here, Vadicherla & Sonawane (2013), the proponents of this system claim that computer based Medical Decision Support System (MDSS) can be useful for the physicians with its fast and accurate decision making process. They opined that predicting the existence of heart disease accurately, results in saving the lives of patients followed by proper treatment. Their objective was to present a MDSS for heart disease classification based on sequential minimal optimization (SMO) technique (which incorporated its features like high accuracy and high speed) in support vector machine (SVM). In using this method, they illustrated the UCI (University College Irvine) machine learning repository data of Cleveland heart disease database and consequently trained the SVM by using SMO technique. Hence, they also claim that given the ease of use and better scaling with the training set size, SMO is a strong candidate for becoming the standard SVM training algorithm. Training a SVM requires the solution of a very large QP (Quantum Platform) optimization problem. SMO algorithm breaks this large optimization problem into small sub-problems. Both the training and testing phases give the accuracy on each record. The results proved that the MDSS is able to carry out heart disease diagnosis accurately in a fast way and it was reported to show good ability of prediction on a large dataset.

2.2 Data Mining in Clinical Decision Support Systems for Diagnosis and Treatment of Heart Disease

According to Amin, Agarwal & Beg (2013) medical errors are both costly and harmful. Medical errors cause thousands of deaths worldwide each year. Hence, a clinical decision support system (CDSS) would offer opportunities to reduce medical errors as well as to improve patient safety. They affirm that one of the most important applications of such systems is in diagnosis and treatment of heart diseases (HD). This is because statistics have shown that heart disease is one of the leading causes of deaths all over the world (CDC Report). Data mining techniques have been very effective in designing clinical support systems because of its ability to discover hidden patterns and relationships in medical data. Here, the proponents also undertook a comparative analysis of the performance and working of six CDSS systems which use different data mining techniques for heart disease diagnosis. They conclude by asserting based on their findings that there is no system to identify treatment options for Heart disease patients. They further claimed that in spite of having a large amount of medical data, it lacked in the quality and the completeness of data thereby creating the need for highly sophisticated data mining techniques to build up an efficient decision support system. They claim that even after doing this, the overall reliability and generalization capability might still be questionable. Hence, the need to build systems which will be accurate, reliable as well as reduce cost of treatment and increase patient care. More so, the building of systems which are understandable and which could enhance human decisions are very germane.

2.3 An Intelligent Decision Support System for the Operating Theater

In 2013 Sperandio, Gomes, Borges, Brito and Almada-Lobo asserted that decision processes inherent in operating theatre organization are often subjected to experimentation, which sometimes lead to far from optimal results. They further affirm that the waiting lists for surgery had always been a societal problem, with governments seeking redress with different management and operational stimulus plans partly due to the fact that the current hospital information systems available in Portuguese public hospitals, lack a decision support system component that could help achieve better planning solutions. As such they developed an intelligent decision support system that allows the centralization and standardization of planning processes which improves the efficiency of the operating theater and tackles the fragile situation of waiting lists for surgery. The intelligence of the system is derived from data mining and optimization techniques, which enhance surgery duration predictions and operating rooms surgery schedules.

2.4 Decision Support System for the Diagnosis of Schizophrenia Spectrum Disorders.

In 2013, Kahn developed a decision support system for the diagnosis of schizophrenia spectrum disorders.

The development of this system is described in four-stages: knowledge acquisition, knowledge organization, the development of a computer-assisted model, and the evaluation of the system's performance. The knowledge is extracted from an expert through open interviews. These interviews aimed at exploring the expert's diagnostic decision making process for the diagnosis of schizophrenia. A graph methodology was employed to identify the elements involved in the reasoning process. Knowledge was first organized and modeled by means of algorithms and then transferred to a computational model created by the covering approach. The performance assessment involved the comparison of the diagnosis of 38 clinical vignettes between an expert and the decision support system. The results showed a relatively low rate of misclassification (18-34%) and a good performance by the decision support system in the diagnosis of schizophrenia, with an accuracy of 66-82%.

3. CLINICAL DECISION SUPPORT SYSTEM (CDSS)

The clinical decision support system is another example of a knowledge based system. A clinical decision support system is an active knowledge system where two or more items of patient data are used to generate case specific recommendations [2].

3.1 Target Area of care

CDSSs assist doctors in assessing various clinical issues from accurate diagnosis of a particular disease to the treatment of the disease. The general target areas of care for CDSS are:

Preventive care which has to do with screening and disease management

Diagnosis which is done based on the patients' signs and symptoms

Follow-up management which has to do with frequent checkups

Hospital Provider Efficiency [8].

3.2 System Design

The system design for CDSS will usually include the following subsystems:
Communication which handles notification and alerts
Knowledge discovery which deals with rules and regulations
Knowledge repository which contains problem solving knowledge [9].

3.3 Factors leading to successful CDSS implementation

The following under listed factors lead to the successful implementation of CDSS:
Simple, user friendly interface
Automated decision support
Timely result
Workflow integration
Continuous Knowledge-base and update support [10].

4. PATTERN CLASSIFICATION METHODS

Pattern classification refers to the theory and algorithms of assigning abstract objects into distinct categories, where these categories are typically known in advance. For this research, the pattern classification methods considered are Decision Trees (DTs), K-Nearest Neighbor (KNN), Naïve bayes Classifier and Support Vector Machine (SVM).

4.1 Decision Trees

A decision tree consists of a root node, branch nodes and leaf nodes. The tree begins with a root node, then further splits into branch nodes and each node represents a choice among various alternatives. The tree then terminates with leaf nodes which are un-split nodes that represent a decision [11]. The classification of decision trees are carried out in two phases:
Tree Building or top down: This is computationally intensive and requires the tree to be recursively partitioned until all the data items belong to the same class.
Tree pruning or bottom top: It is conducted to improve the prediction and classification of the algorithm and minimize the effects of over-fitting which may lead to misclassification of errors [12].
Some notable decision tree algorithms include Classification and Regression Trees (CART), Iterative Dichotomiser 3 (ID3), C4.5 and C5.0.
The advantages of decision trees include:

- They are easy to interpret and comprehend
- They can handle both metric and non-metric data as well as missing values which are frequently encountered in clinical studies.
- Little data preparation is required since data does not need to be normalized.
- They can handle data in a short time frame.
- They can be developed using common statistical techniques.

The disadvantages associated with decision trees include:

- They can over fit the data and create complex trees that may not generalize well.
- A small change in the size of a dataset could result in a completely different tree

4.2 K-Nearest Neighbor

K-Nearest Neighbor (k-NN) is instance based learning for classifying objects based on closest training examples in the feature space. It is a type of lazy learning where the function is only approximated locally and all computations are deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. If k=1, then the object is simply assigned to the class of its nearest neighbor. The k-NN algorithm uses all labeled training instances as a model of the target function. During the classification phase, k-NN uses a similarity-based search strategy to determine a locally optimal hypothesis function. Test instances are compared to the stored instances and are assigned the same class label as the k most similar stored instances.

4.3 Bayes Classifier

A Bayesian network is a model that encodes probabilistic relationships among variables of interest. This technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data. However, a serious disadvantage of using Bayesian networks is that their results are similar to those derived from threshold-based systems, while considerably higher computational effort is required.

4.4 Support Vector Machine

Support Vector Machines have been proposed as a novel technique for intrusion detection. An SVM maps input (real-valued) feature vectors into a higher-dimensional feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization. Structural risk minimization seeks to find a hypothesis (h) for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to a quadratic optimization problem. The implementation of SVM intrusion detection system has two phases which are training and testing. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification.

Table 1. Comparison of Some Pattern Classification Algorithms (Source: Patel et al, 2012)

Classifier	Method	Parameters	Advantages	Disadvantages
Support Vector Machine	A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional	The effectiveness of SVM lies in the selection of kernel and soft margin parameters. For kernels, different pairs of	1. Highly Accurate 2. Able to model complex nonlinear decision boundaries 3. Less prone to over fitting than other methods	1. High algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale

	space, which can be used for classification, regression or other tasks.	(C, γ) values are tried and the one with the best cross-validation accuracy is picked. Trying exponentially growing sequences of C is a practical method to identify good parameters.		tasks. 2. The choice of the kernel is difficult 3. The speed both in training and testing is slow.
K Nearest Neighbour	An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its *k* nearest neighbours (*k* is a positive integer). If *k* = 1, then the object is simply assigned to the class of its nearest neighbour.	Two parameters are considered to optimize the performance of the kNN, the number k of nearest neighbour and the feature space transformation.	1. Analytically tractable. 2. Simple in implementation 3. Uses local information, which can yield highly adaptive behaviour 4. Lends itself very easily to parallel implementations	1. Large storage requirements. 2. Highly susceptible to the curse of dimensionality. 3. Slow in classifying test tuples.
Bayesian Method	Based on the rule, using the joint probabilities of sample observations and classes, the	In Bayes, all model parameters (*i.e.*, class priors and feature probability distributions) can be	1. Naïve Bayesian classifier simplifies the computations. 2. Exhibit high accuracy and speed	1 The assumptions made in class conditional independence. 2. Lack of available probabilit
	algorithm attempts to estimate the conditional probabilities of classes given an observation.	approximated with relative frequencies from the training set.	when applied to large databases.	y data.
Decision Tree	Decision tree builds a binary classification tree. Each node corresponds to a binary predicate on one attribute; one branch corresponds to the positive instances of the predicate and the other to the negative instances.	Decision Tree Induction uses parameters like a set of candidate attributes and an attribute selection method.	1. Construction does not require any domain knowledge. 2. Can handle high dimensional data. 3. Representation is easy to understand. 4. Able to process both numerical and categorical data.	1. Output attribute must be categorical. 2. Limited to one output attribute. 3. Decision tree algorithms are unstable. 4. Trees created from numeric datasets can be complex.

5. METHODOLOGY

A very comprehensive dataset (Velvet, 2008) consisting of 100,000 instances compiled from the UCI (University of California, Irvine) data repository was used. This dataset was translated into the Attribute Relational File Format (ARFF) which is one of the file formats recognized by the WEKA (Waikato Environment for Knowledge Analysis) software in which the distinct genotypic attributes used for this work were highlighted.

The dataset was then induced with Classification algorithms namely C4.5 decision trees, Support Vector Machine (SVM), K-Nearest neighbor algorithm and Bayes Classifier Algorithm. The Classification algorithms were evaluated using the Waikato Environment for Knowledge Analysis software version 3.6.7 based on the percentage of correctly classified instances with the C4.5 decision trees having 61.0734%, the Support Vector Machine (SVM) algorithm had 50.0515%, the Bayes Classifier Algorithm had 50.2045% and the K-Nearest Neighbor algorithm had 50.1235%. Sequel to the result obtained from this evaluation, the C4.5 decision trees turn out as the Classification algorithm with the highest accuracy for this research. Thereafter, a decision tree program was written in Java with 38 lines of code for the core program

to implement the C4.5 decision tree algorithm that will provide the requisite intelligence for this Clinical decision support system and help it make the right decisions promptly when supplied with patient information. The C4.5 decision tree algorithm will be embedded in the classification/prediction algorithm section of the clinical decision support system.

6. CONCLUSION AND RECOMMENDATION

This research work finds its significance in all parts of the world where people live with the health challenge caused by the hepatitis B virus, thus it is very germane as it provides a sort of panacea to the eventual development of the condition known as hepatitis B for people who are susceptible to the condition, hence they can be aware of their susceptibility ahead of time and can be able to take the necessary precautionary measures to forestall their development of the illness, thus saving them from the trauma they would have inevitably suffered.

The research is a milestone in the sub-field of health informatics as it provides a readily available Clinical Decision Support System to serve as a reliable assistant to the medical practitioners that are more often than not burdened by the overwhelming and seemingly intimidating number of patients they need to attend to routinely. This has culminated in a lot of fatal errors on the part of the medical practitioners which has led to the loss of innocent lives hence, the introduction, consequent adoption and deployment of this Knowledge Based Intelligent Clinical Decision Support System for the prediction of hepatitis B becomes expedient especially in the third world countries, the vast majority of who lag behind in terms of technological innovations and advancement and as a result are alien to the terrific results gotten from the use of these clinical decision support systems.

For further work another enthusiastic researcher can go a step further in this work by introducing other highly efficacious algorithms that can be used alongside the C4.5 decision tree algorithm used in this work, so as to have a hybrid system that will take decisions faster and generate more accurate decisions than those that will be given by the proposed system.

7. REFERENCES

[1] Power, D.J. (1999). Decision Support Systems Glossary.http://DSSResources.COM/glossary

[2]Chen, J.Q & Lee, S.M. (2002). An exploratory cognitive DSS for strategy decision making. *Elsevier Science B.V.*

[3] Naumann, H, Scott R.M, Snitbhan R, Bancroft W.H, Alter H.J & Tingpalapong, M (2006). Experimental transmission of hepatitis B virus by semen and saliva. *International Journal of Infectious Diseases*, 23(8), 27-35.

[4] Vadicherla, D. & Sonawane, S. (2013). Decision support system for heart disease based on sequential minimal optimization in support vector machine. *International Journal of Engineering Sciences & Emerging Technologies*, 4(2), 19-26.

[5] Amin, S.U, Agarwal, K & Beg R. (2013). Data mining in clinical decision support system for diagnosis, prediction and treatment of heart disease. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 2(1), 56-67.

[6] Sperandio F, Gomes C, Borges J, Brito A.C &Almada-Lobo B. (2013). An intelligent Decision Support System for the Operating Theatre: A Case Study. *Automation Science and Engineering, IEEE Transactions on Robotics & Control Systems*, 99.

[7] Kahn, R., Perkins, D., Lieberman, J.(2012). Predictors of treatment response in patients with first-episode prostate cancer disorder. *The British Journal of Gynecology*, 185(1),18-24.

[8] Berner, E.S. (2009). Clinical Decision Support Systems: State of Art. Rockville: AHRQ Publication 12(5), 90-134.

[9] Frize, M. (2005). Conceptual Framework of Knowledge Management for Ethical Decision making Support in Neonatal Intensive Care. *IEEE Transactions of Information Technology in Biomedicine*, 9(7), 205-215.

[10] Peleg, M., & Tu, S. (2011). Decision Support, Knowledge Representation and Management in Medicine. Stanford Centre for Biomedical Informatics Research. Last accessed: December 17, 2011 from http://bmir.stanford.edu/file_asset/index.php/1009/SMI-2006-1088.pdf

[11] Peng, W., Chen J. & Zhou H. (2006). An Implementation of ID3-Decision Tree Learning Algorithm. University of New South Wales. Last accessed: December17, 2011 from http://web.arch.usyd.edu.au/~wpeng/DecisionTree2.pdf.

[12] Anyanwu, M.N., & Shiva, A.G., (2009). Comparative Analysis of serial decision tree classification algorithms. *International Journal of Computer Science and Security*, 3(3), 230-240.

[13] Neth, Q, & Alter M (2006). Epidemiology of hepatitis B in Europe and worldwide. *International Journal of Hepatology*, 39(4), 66-69.

[14] Buster, R, & Hyams K.C (2006). Risks of chronicity following acute hepatitis B virus infection. *International Journal of Clinical Infectious Diseases*, 20(3), 992–1000.

[15] Hepatol, O.M, Orito, E, Mizokami, M, Sakugawa, H, Michitaka, K, Ishikawa, K & Ichida, T (2003). A case-control study for clinical and molecular biological differences between hepatitis B viruses of genotype B and C. *Canadian Journal of Hepatology*, 14(7), 908-921