The Impact of Socio-demographic Factors on the Prevalence of Hepatitis B in Sokoto, Nigeria: An Association Rule Mining Approach

Rufai Ahmad
Department of Computer Science
Sokoto State University, Nigeria

Fatimah Jumare
Department of Microbiology
Sokoto State University, Nigeria

Mahmood Umar
Department of Computer Science
Sokoto State University, Nigeria

ABSTRACT

The hepatitis B virus (HBV) is a dangerous liver infection that results in hepatitis. Despite efforts to create awareness of the disease and the development of vaccines to prevent people from being infected, the prevalence of HBV in developing countries is still very high. This study aimed to apply apriori algorithm to investigate the relationship between socio-demographic factors and HBV status among patients visiting various hospitals within the metropolis of Sokoto state, Nigeria. Using data from 423 patients, we found that younger participants aged 26-35 have the highest positive cases. The rules generated from the apriori algorithm suggest low awareness of both HBV and the HBV vaccines. However, having some knowledge about HBV and the vaccine tends to be associated with negative HBV status. These findings have implications for healthcare bodies in that they could inform how to devise strategies for treatment and awareness campaigns to reduce the spread of the disease.

General Terms

Data Mining, Apriori Algorithm, Hepatitis B

Keywords

HBV Prevalence, Data Mining, Association Rules

1. INTRODUCTION

Hepatitis B (HBV) is a viral disease that can attack the liver and cause acute and chronic infection [1]. Approximately 296 million people worldwide carry HBV [1]. The latest estimate shows that 820,000 died from complications resulting from HBV [2]. In Nigeria, the most recent estimation of those living with the virus is 24 million [3]. Several studies have confirmed the relationship between socio-demographic characteristics, such as gender, age, marital status, and the prevalence of HBV. For example, [4] found a connection between the marital status of women in Eretria and their HBV status. In addition, the authors also found a connection between women's formal education and their hepatitis status.

Similarly, authors in [5] used Chi-squared tests to compare socio-demographic characteristics among 329 HBV positive Vietnamese immigrants in the US. The authors found that participants under 30 years had the highest HBV prevalence, followed by 41-50 years of age. In Nigeria, authors in [6] reviewed published studies and unpublished theses on HBV in Nigeria conducted at Ahmadu Bello University, Zaria, from 2000-2014. The results of this analysis confirmed a prevalence of 3.9 to 50.7%. Further analysis also showed that HBV Infection was more prevalent in males than females, with infection peak in the age group 20 to 30 years [6]. Another study by [7] suggests that economic and social factors are connected to the prevalence of HBV in Turkey. Specifically, the authors found that age, gender, migration, education, awareness, social welfare, and occupation were significant factors in determining the prevalence of HBV in Turkey.

A 2020 analysis of the educational gap in Nigeria revealed a wide education gap between most states in the Northern part of Nigeria and their southern counterparts [8]. According to the 2017 data from the country's national statistics body, the states with the highest number of people who can neither read nor write are located in the Northern region [9]. This is worrisome, considering findings showing the relationship between education and health [10]. Furthermore, the literature shows that economic and social factors are key to HBV prevalence [7]. Sokoto is a state in the Northern part of Nigeria and, therefore, can be used as a case study for investigating the relationship between the prevalence of HBV and economic and social factors.

Data mining is the process of identifying valid, novel, and useful patterns from an existing dataset [11]. Data mining can help organisations to find those relationships and patterns that may otherwise hide in large datasets. The powerful nature of data mining techniques and their ability to mine interesting patterns from large datasets has seen their application in healthcare [12]. Data mining functionalities include classification, clustering, prediction, and association. Association rule mining is one of the most important applications of data mining. The technique was first introduced in 1993 and is used to find relationships between a set of items in a dataset [13].

This study used an association rule mining algorithm in particular, the apriori algorithm, to analyze medical data. Specifically, the research aims to investigate the relationship between socio-demographic factors and HBV status among patients visiting various hospitals within the metropolis of Sokoto state, Nigeria.

The rest of the paper is organised as follows. Section 2 describes the background and related work, and the material and methods are described in Section 3. The results of this study are provided in section 4. Section 5 discusses the findings and concludes the study.

2. BACKGROUND AND RELATED WORK

2.1 Association Rule Mining and Apriori Algorithm

Association rule mining is a data mining technique for finding patterns in a dataset. The apriori algorithm is a popular algorithm based on market basket analysis used to find frequent patterns and association rules in a dataset [14]. The algorithm was first introduced by R.Agrawal et al in 1993. Apriori algorithm uses an iterative approach to find individual frequent items in a database and extend them to larger itemsets. The frequent itemsets are a set of itemset that frequently occur together. Apriori algorithm helps find trends in data by revealing the association rules that highlight relationships between different itemset.

The following outlined how Apriori algorithm finds rules within a dataset.

1. **Frequent Itemset**: In this stage, the algorithm identifies individual items in the dataset and their frequencies of occurrences.

2. **Candidate Itemset generation**: The algorithm generates candidate itemsets from the frequent itemset identified in the previous stage. For example, if **X and Y** are frequent then {**X , Y** } is candidate itemset.

3. **Find the candidate itemset in the dataset**: The algorithm scans the given dataset to count the occurrences of the candidate itemset identified in the previous stage. The algorithm also decides which candidate Itemsets move to the next stage based on criteria known as minimum support.

4. **Generate association rules**: The algorithm generates association rules from the frequent itemset that satisfy the minimum confidence threshold. The following approach is used to generate the association rules:

   Let **I**={....} be a set of ‘n’ binary attributes called items.

   Let **D**= {...} be a set of transactions called a database.

Each transaction in **D** has a unique transaction ID and contains a subset of the items in **I**. A rule is defined as the implication of the form **X** **Y**→where **X**, **Y** are subsets of **I and X∩Y=ϕ.** The set of items X and Y are called antecedent and consequent of the rule [15].

In a nutshell, An **association rule in apriori algorithm** is of the form *X* => *Y*, where *X* = {*x*1, *x*2, ..., *xn*}, and *Y* = {*y*1, *y*2,..., *ym*} are sets of items, with *xi* and *yj* being distinct items for all *i* and all *j*. This association states that if a customer buys *X*, he or she is also likely to buy *Y*. In the context of medical research, this could be reframed as if a patient has a symptom X, then he/she is likely to have a disease Y.

The Apriori algorithm uses the following measures of significance to select interesting rules from a set of all possible rules.

Support: The proportion of transactions in the dataset which contain the itemset; i.e. if we have the rule **X & Y =Z**, support is the probability that a transaction contains {**X, Y, Z**} [14].

Confidence: Is the percentage of cases to which all the rules apply (i.e. if we have the rule ***X*** => ***Y***, confidence is the probability that a record having ***X*** also contain ***Y***)

2.2 Related Work

Prior studies have demonstrated how the apriori algorithm could be applied to medical data to discover important hidden patterns. For example, [16] used the apriori algorithm to identify relationships between procedures performed on a patient and the reported diagnoses from medical data containing 30 million line items[16]. The researchers obtained association rules that show relationships between medical procedures and the corresponding diagnoses. The author in [17] presented an improved version of the apriori algorithm to find the relationship between breast cancer recurrences and other attributes using SQL Server 2005 Analysis Services. More recently, [18] applied the apriori algorithm to investigate the combinations of care needs for people living with dementia and their caregivers based on subtypes of dementia.

Nevertheless, while all these studies have focused on finding associations and relationships between symptoms and diseases, none among them and within the literature has focused on finding the association between different biological, social and environmental factors on the prevalence of HBV. As a result, this study aims to fill this gap using data from Nigeria.

3. MATERIALS AND METHOD

3.1 Study Design

The study was cross-sectional and descriptive, and it was designed to determine the relationship between HBV surface antigen among participants and socio-demographic factors. The study used a structured questionnaire to collect various socio-demographic measurements from participants, including gender, age, occupation, level of education, locality, marital status, knowledge of HBV vaccine, HBV vaccine status, and HBV status. Data was collected from patients attending different hospitals in Sokoto metropolis, including Noma Children Hospital, Specialist Hospital, Women, Children Welfare Clinic, and Maryam Abacha Women and Children Hospital. Ethical approval was received from the Sokoto State Ministry of Health (SMH/1580/V.IV) before conducting the study. The questionnaire was in English but translated into Hausa, the native language most of the residents in Sokoto understood.

3.2 Study Population

Data from 423 patients attending HBV surface antigen (HBsAg) screening at various hospitals from 20 September 2021 to 7 January 2021 was collected. Most of the participants were attending the screening following a request by their GP. The participants were first interviewed, and after consenting to take part in the research, their socio-demographic measurements were collected. A blood sample was then collected from each participant to determine their HBsAg status. For each participant, the following data was collected: gender, age, occupation, level of education, marital status, area/locality, knowledge of hepatitis b, knowledge of hepatitis b vaccine, hepatitis b vaccine status, and hepatitis b status(see also Table 1).

3.3 Viral Diagnosis

One millilitre (1ml) of Blood sample was collected through a septic venin puncture for HBV serology. An assay for HBsAg was done using an Enzyme-linked immunosorbent assay (ELISA) kit produced by Abbatt to determine HBs Ag, which has a 99.9 and 99.0% sensitivity, respectively. The test was then read after a waiting interval of 15 minutes to 24 minutes to 24 hours (as specified by the manufacturer).

3.4 Data Preprocessing and Analysis Strategy

The data was transferred from the paper-based questionnaire to SPSS statistical software for cleaning and analysis. Descriptive statistics such as frequency and percentages for all the variables of interest were calculated. The number of participants was reduced to 419 after removing 4 cases with missing data. These cases were removed because the participants didn’t report their Hepatitis B status.

To investigate the relationship between socio-demographic factors and HBV status, the WEKA data mining software was used. This software has a collection of machine-learning algorithms for data mining tasks [19]. We used the Apriori algorithm module of WEKA to perform association rule mining of the dataset.

Table 1: Attributes with their corresponding values

Attribute	Description	Values
Gender	Gender	(Male, Female)
Age_Group	Participants Age Group	(18-25, 26-35, 36-45, 46-55, 56-65, 66+)
Occupation	Occupation	(business, civil servant, employed, farming, housewife, retired)
Education	Education	(college, junior, none, primary, secondary, university)
Marital Status	Participants Marital Status	(divorce, married, single, widow)
Area/Locality	Participants residential area	(rural, urban)
Knowledge_about_HepB	Knowledge about hepatitis	(Yes, No)
HepB_Vacc_knowledge	Participants knowledge about HBV vaccine	(Yes, No)
Vaccinated	Hepatitis B vaccination Status	(no,' I don't know',yes)
HepB_Status	Hepatitis B Status	(Negative, Positive)

4. RESULTS

From the 419 cases, the results revealed that 43.9% (n=184) were positive HBV cases and 56.1 (n=235) were negative (See Figure 1). The male gender has the highest HBV-positive cases (n=120). Age group 26-35 has the highest number of positive cases (n=80), followed by 18-25 (n=54), and 36-45 (n=37). Across the educational level, the results show that those with college-level education (n=74) have the highest number of positive cases, followed by those with no qualification (n=51), then junior secondary school (n=21). Married participants have the highest positive cases (n=142). Analysis of the area/locality variables shows that (n=91) of participants live in rural areas while (n=93) live in urban. Considering that the total number of positive and negative cases in rural areas is (n=159), positive cases in rural areas are still high. Participants who indicated that they were unaware of the HBV vaccine have the highest number of positive cases (n=145) compared to those who are aware of it (n=39).

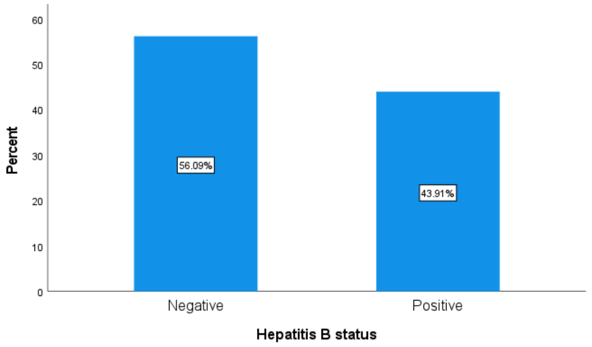

Fig 1: Participants Hepatitis B Status

To investigate the relationship between the socio-demographic factors used in this study and the participants' HBV status, this study used a three-stage approach. In the first stage, the algorithm was run on the complete dataset (i.e. both positive and negative cases). Similar to [15]the minimum support to 0.1 and confidence to 0.9, respectively, whilst living the number of rules to be returned to default, currently set at 10 in Weka. The results from this run returned the five best rules linking the participants' HBV status and socio-demographic variables, as seen in Table 2. This table shows that most of the positive cases are male, have gone to college and are unaware that there is a vaccination for HBV. The table's last rule suggests that those aware of the HBV vaccine tested negative for HBV.

Table 2 Rules linking the participants' Hepatitis B status and socio-demographic variables

Rule No.	Antecedent	Consequent	Confidence %
1	Gender=male , Level_of_Education=college , Marital_Status= married , Vaccinated=no	HepB Status=Positive	93
2	Gender=male , Level_of_Education=college , HepB_Vacc_knowledge=no , Vaccinated=no	HepB Status=Positive	92
3	Gender=male , Level_of_Education=college , Marital_Status= married , HepB_Vacc_knowledge=no	HepB Status=Positive	91
4	Gender=male, Level_of_Education=college , HepB_Vacc_knowledge=no	HepB Status=Positive	91
5	Marital_Status= single, HepB_Vacc knowledge=yes	HepB_Status= Negative	90

In the next analysis stage, the data was filtered according to HBV status (i.e. positive and negative cases). Thereafter, the analysis was performed on each group. The aim was to understand what factors were associated with each category—for example, the socio-demographic factors associated with being free of HBV. Starting with the positive cases, Table 3 suggests that similar to what we found in Table 2, most of the

HBV were male, married, and unaware that they could get a vaccine for HBV. For example, rule 10 shows that those who were married, unaware of the existence of a vaccine for HBV and not vaccinated tend to be positive for HBV.

Table 3: Best rules in HBV-positive cases

Rule No.	Antecedent	Consequent	Confidence %
1	Vaccinated=no	HepB Status=Positive	100
2	HepB_Vacc _knowledge =no	HepB Status=Positive	100
3	Marital_Status= married	HepB Status=Positive	100
4	HepB_Vacc _knowledge =no, Vaccinated=no	HepB Status=Positive	100
5	Gender=male	HepB_Status= Positive	100
6	Marital_Status= married , Vaccinated=no	HepB_Status= Positive	100
7	Marital_Status= married, HepB_Vacc _knowledge =no,	HepB_Status = positive	100
8	Gender=male, HepB_Vacc _knowledge =no	HepB_Status= Positive	100
9	Gender=male, Vaccinated=no	HepB_Status=Positive	100
10	Marital_Status= married, HepB_Vacc _knowledge =no, Vaccinated=no	HepB_Status= Positive	100

Table 4 shows the results from the analysis of those with negative HBV. Similar to the positive cases, the rules suggest that most participants with negative cases resided in the urban area and were also married. Interestingly, rule 3 shows that there were unvaccinated participants who were negative for HBV at the time of the survey. The most interesting rules in Table 4 are rules 7-10. Rule 7 shows that those who had some knowledge about what HBV is and were aware of the vaccine for HBV tend to be free from the virus. Rule 8 shows that those who knew about HBV tend to also know about the vaccine for HBV. Rule 9 and 10 also show a similar trend to rule 7, where knowledge about HBV and the vaccine for HBV tend to be associated with negative HBV status.

Table 4: Best rules in HBV-negative cases

Rule No.	Antecedent	Consequent	Confidence %
1	AreaLocality=urban	HepB status =Negative	100
2	Marital_Status= married	HepB Status= Negative	100
3	Vaccinated=no	HepB status =Negative	100
4	Gender=female	HepB status = Negative	100
5	Knowledge_about_HepB =yes	HepB_Status = Negative	100
6	HepB_Vacc _knowledge =yes	HepB_Status= Negative	100
7	Knowledge_about_HepB=yes, HepB_Vacc _knowledge =yes	HepB_Status = Negative	100
8	HepB_Vacc _knowledge =yes	Knowledge_about _HepB =yes	98
9	HepB_Vacc_knowledge =yes, HepB_Status= Negative	Knowledge_about _HepB=yes	98
10	HepB_Vacc_knowledge =yes	Knowledge_about _HepB =yes, HepB_Status=Negative	98

5. DISCUSSION

This study investigated the relationship between socio-demographic factors and HBV status in Sokoto state, Nigeria. Our findings show that participants aged 26-35 have the highest positive cases in the study's sample. These findings agree with the results from [6], where a high prevalence of HBV infection within the age group 20-30 years was found. Furthermore, considering that Nigeria only began immunizing newborn babies in 2004 [20], the prevalence of HBV within the age group 26-35 is likely due to the transmission of HBV from mothers to their children [21]. In addition, this study also shows more positive cases in rural areas than in urban areas. Our findings agree with a recent meta-analysis of data published between 2010-2019 [22]. The prevalence of HBV in rural areas may suggest a lack of awareness about the implications of HBV among rural dwellers. The rules generated from the apriori algorithm show a prevalence of HBV in male participants who attended college and were unaware that there was a vaccination for HBV. These rules suggest that the selected sample may have low awareness of HBV. This low awareness may have contributed to the prevalence of the virus in the selected communities. Further supporting this argument, it was also found that those who had some knowledge about HBV and were aware of the vaccine for HBV tend to be free from the virus. These results suggest the need for healthcare bodies to create more awareness about the diseases and the vaccinations available. Furthermore, concerned bodies should create awareness initiatives that target the rural areas of the state.

6. CONCLUSION

Most previous studies on the relationship between socio-demographic factors and HBV status rely on traditional conventional statistical methods such as the Pearson correlation Chi-square test, which rely on assumptions about data and its distributions. In this study, an association rule mining algorithm was used to investigate the association between socio-demographic factors and HBV status. The rules generated from the algorithm show that those aware of the HBV vaccine end up testing negative for HBV. Furthermore, those unaware of the existence of a vaccine for HBV and who are not vaccinated tend to be positive for HBV. These findings have broader implications. For example, they could inform how healthcare bodies devise strategies for treatment and awareness campaigns to reduce the spread of the disease. This research's novel contribution is that it provides insight into the prevalence of HBV in Sokoto state, Nigeria. In addition, it also provides a new dataset that other researchers could use to understand the relationship between socio-demographic factors and HBV prevalence.

7. ACKNOWLEDGEMENT

We would like to appreciate Nigeria's Tertiary Education Trust Fund (TETFUND) for providing the grant to carry out this study. We would also like to thank the Sokoto State Ministry of Health for giving ethical approval to carry out this study.

8. REFERENCES

[1] WHO, "Hepatitis B," 2023. https://www.who.int/news-room/fact-sheets/detail/hepatitis-b#:~:text=WHO estimates that 296 million,carcinoma (primary liver cancer). (accessed Oct. 07, 2023).

[2] Hepatitis B Foundation, "Hepatitis B Facts and Figures," 2022. https://www.hepb.org/what-is-hepatitis-b/what-is-hepb/facts-and-figures/ (accessed Jan. 15, 2022).

[3] Punch, "World Hepatitis Day 2021: Nigeria dallies as world races to end disease by 2030," 2021. https://www.premiumtimesng.com/news/headlines/476145-world-hepatitis-day-2021-nigeria-dallies-as-world-races-to-end-disease-by-2030.html (accessed Jan. 15, 2022).

[4] N. Fessehaye, A. Berhane, and H. Ahimed, "Prevalence of hepatitis B virus infection and associated seromarkers among pregnant women in Eritrea," *J. Hum. Virol. Retrovirology*, vol. 6, no. 1, pp. 30–38, 2018.

[5] C. Strong, K. Hur, F. Kim, J. Pan, S. Tran, and H.-S. Juon, "Sociodemographic characteristics, knowledge and prevalence of viral hepatitis infection among Vietnamese Americans at community screenings," *J. Immigr. Minor. Heal.*, vol. 17, no. 1, pp. 298–301, 2015.

[6] M. Aminu, "HEPATITIS B INFECTION IN NIGERIA: A REVIEW," 2014.

[7] S. Tosun, O. Aygün, H. Ö. Özdemir, E. Korkmaz, and D. Özdemir, "The impact of economic and social factors on the prevalence of hepatitis B in Turkey," *BMC Public Health*, vol. 18, pp. 1–9, 2018.

[8] Daily Trust, "Why North Fails To Close Education Gap With South," 2020. https://dailytrust.com/why-north-fails-to-close-education-gap-with-south (accessed Jul. 08, 2021).

[9] A. Amzat, "Despite decades of funding, literacy level in the northern states remains low," 2017. .

[10] L. Feinstein, R. Sabates, T. M. Anderson, A. Sorhaindo, and C. Hammond, "What are the effects of education on health," in *Measuring the effects of education on health and civic engagement: Proceedings of the Copenhagen symposium*, 2006, pp. 171–354.

[11] H. M. Chung and P. Gray, "Data mining," *J. Manag. Inf. Syst.*, vol. 16, no. 1, pp. 11–16, 1999.

[12] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002.

[13] M. Ilayaraja and T. Meyyappan, "Mining medical data to identify frequent diseases using Apriori algorithm," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013, pp. 194–199.

[14] M. Karaolis, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 6238–6241.

[15] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, pp. 447–459, 2018.

[16] A. M. Doddi SS Ravi, David C. Torney, Srinivas, "Discovery of association rules in medical data," *Med. Inform. Internet Med.*, vol. 26, no. 1, pp. 25–33, 2001.

[17] R. Hu, "Medical data mining based on association rules," *Comput. Inf. Sci.*, vol. 3, no. 4, p. 104, 2010.

[18] K.-M. Jhang, M.-C. Chang, T.-Y. Lo, C.-W. Lin, W.-F. Wang, and H.-H. Wu, "Using the Apriori algorithm to classify the care needs of patients with different types of dementia," *Patient Prefer. Adherence*, pp. 1899–1912, 2019.

[19] U. of Waikato, "Weka 3: Machine Learning Software in Java," 2022. https://www.cs.waikato.ac.nz/ml/weka/ (accessed Jan. 16, 2022).

[20] A. E. Sadoh and A. Ofili, "Hepatitis B infection among Nigerian children admitted to a children's emergency room," *Afr. Health Sci.*, vol. 14, no. 2, pp. 377–383, 2014.

[21] D. Ndububa *et al.*, "Prospective cohort study of prevention of mother to child transmission of hepatitis B infection and 9 months follow-up of hepatitis B-exposed infants at Ile-Ife, Nigeria," *BMJ Open*, vol. 12, no. 11, p. e063482, 2022.

[22] B. I. Ajuwon, I. Yujuico, K. Roper, A. Richardson, M. Sheel, and B. A. Lidbury, "Hepatitis B virus infection in Nigeria: a systematic review and meta-analysis of data published between 2010 and 2019," *BMC Infect. Dis.*, vol. 21, pp. 1–15, 2021.