









One-year mortality prediction of patients with hepatitis in Kazakhstan based on administrative health data: A machine learning approach

Iliyar Arupzhanov¹ , Dmitry Sysoyev² , Aidar Alimbayev³ , Gulnur Zhakhina² , Yesbolat Sakko² ,
Sauran Yerdessov² , Amin Zollanvari¹ , Abduzhappar Gaipov^{2*} 

¹School of Engineering and Digital Sciences, Nazarbayev University, Astana, KAZAKHSTAN

²Department of Medicine, School of Medicine, Nazarbayev University, Astana, KAZAKHSTAN

³Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

*Corresponding Author: abduzhappar.gaipov@nu.edu.kz

Citation: Arupzhanov I, Sysoyev D, Alimbayev A, Zhakhina G, Sakko Y, Yerdessov S, Zollanvari A, Gaipov A. One-year mortality prediction of patients with hepatitis in Kazakhstan based on administrative health data: A machine learning approach. *Electron J Gen Med.* 2024;21(6):em618. <https://doi.org/10.29333/ejgm/15747>

ARTICLE INFO

Received: 14 Oct. 2024

Accepted: 16 Dec. 2024

ABSTRACT

Background and objective: Hepatitis B virus (HBV) and hepatitis C virus (HCV) are major contributors to chronic viral hepatitis (CVH), leading to significant global health mortality. This study aims to predict the one-year mortality in patients with CVH using their demographics and health records.

Methods: Clinical data from 82,700 CVH patients diagnosed with HBV or HCV between January 2014 and December 2019 was analyzed. We developed a machine learning (ML) platform based on six broad categories including linear, nearest neighbors, discriminant analysis, support vector machine, naïve Bayes, and ensemble (gradient boosting, AdaBoost, and random forest) models to predict the one-year mortality. Feature importance analysis was performed by computing SHapley Additive exPlanations (SHAP).

Results: The models achieved an area under the curve between 0.74 and 0.8 on independent test sets. Key predictors of mortality were age, sex, hepatitis type, and ethnicity.

Conclusion: ML with administrative health data can be utilized to accurately predict one-year mortality in CVH patients. Future integration with detailed laboratory and medical history data could further enhance model performance.

Keywords: chronic viral hepatitis, mortality prediction, machine learning, SHAP analysis, artificial intelligence

INTRODUCTION

Millions of people are impacted by chronic viral hepatitis (CVH), which makes it a substantial healthcare challenge around the globe. It is characterized by persistent inflammation of the liver caused by viral infections, primarily the viruses known as hepatitis B virus (HBV) [1] and hepatitis C virus (HCV) [2]. According to the World Health Organization (WHO) report [3], there were an estimated 296 million people who had chronic hepatitis B, while nearly 60 million individuals had chronic hepatitis C (CHC), which made up 3.8% and 0.8% of the world population, respectively in 2019. In Kazakhstan [4], the inpatient and outpatient registries recorded a total of 82,700 individuals diagnosed with HBV or HCV between 2014 and 2019.

Infections caused by HBV [1] and HCV [2] are the primary factors leading to chronic cirrhosis, hepatocellular carcinoma, liver failure, and other liver-related deaths. WHO reported that in 2019, HBV and HCV infections caused approximately 1.1 million deaths [3]. In the absence of further interventions, it is projected that the estimated death toll from hepatitis could

reach 19 million from 2015 to 2030. WHO has set a target to reduce mortality rates by 65% by 2030 [5]. Therefore, it is important to develop an effective mortality prediction system to assist clinicians in tailoring treatment strategies and enhancing the survival rates of HBV and HCV patients.

Machine learning (ML) models have been widely utilized in various healthcare applications. Several research papers have employed ML techniques for predicting hepatitis [6-8]. For instance, it was used six ML algorithms, including logistic regression (LR), K-nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), XGBoost (XGB), and artificial neural networks (ANN) to predict CHC [8]. Moreover, ML techniques have been applied to predict treatment response in patients with CVH [9-11]. It was predicted the treatment response against L-ornithine L-Aspartate medicine in hepatitis C patients utilizing various ML techniques, including naïve bayes (NB), random forest (RF), DT, and KNN, to name a few [10]. Additionally, ML algorithms were used to diagnose the stage of hepatitis [12]. Researchers have also leveraged ML models to accurately predict the risk of mortality in patients diagnosed with CVH, utilizing clinical and administrative data [13-15]. However, there were no studies focusing on the

Table 1. Description of clinical variables used in yearly-specific cohorts

Feature	Description	Unit	Type
Type of hepatitis	Chronic hepatitis C or chronic hepatitis B without delta function	Binary	Categorical
Sex	Female or male	Binary	Categorical
Age	Age at the diagnosis of hepatitis	Years	Numeric
Ethnicity	Kazakhs, Russians, and others	Ternary	Categorical
Cirrhosis	Complication for hepatitis (yes /no)	Binary	Categorical
Duration of hepatitis	Time from initial diagnosis to December 31 st of the year preceding the prediction	Years	Numeric
Hospitalization	Whether the patient was hospitalized or not (yes/no)	Binary	Categorical

predicting one-year mortality for CVH patients using only administrative data, which encompasses demographic data (age and sex), comorbidities and complications, diagnoses, and characteristics of service providers. This focus is warranted because these information are generally easy and inexpensive to collect.

To address this issue, we developed an ML platform to build a model that predicts one-year mortality in patients with CVH. The ML platform was developed using the clinical data of a group of CVH patients diagnosed with hepatitis between January 2014 to December 2019, collected from the Kazakhstan unified national electronic health system (UNEHS) [16]. The dataset was split into four groups to predict the one-year mortality, using clinical data gathered until the end of the previous year. Our study demonstrates the feasibility and robustness of this ML platform, which utilizes aggregated nationwide administrative healthcare data to predict one-year mortality in CVH patients of Kazakhstan. Additionally, we identified and ranked the clinical variables in the developed predictive models.

A one-year mortality prediction model for hepatitis patients can be used as an auxiliary tool in clinical practice. It would enable medical professionals to create personalized treatment strategies and take preventive measures to reduce negative outcomes. Additionally, this model would help in better managing healthcare resources, highlighting the need for regular monitoring or additional care for patients considered to be at higher risk.

RESULTS

Data Description

This study aims to use administrative health data to predict the one-year mortality of CVH patients. To accomplish this goal, clinical records of individuals diagnosed with either HBV or HCV were extracted from the UNEHS [4] database between January 2014 and December 2019 (for details on how the patient cohort was selected, refer to the materials and methods section). Patients with missing vital outcomes, either deceased or alive, were excluded from the analysis. The remaining data were then divided into four sub-cohorts, corresponding to the years 2016, 2017, 2018, and 2019, to predict yearly mortality for each year using clinical information available until the end of the previous year. For instance, the 2017-cohort was formed from the patients who were alive as of 31st December 2016 with known outcome variable (for further details, refer to [Appendix A](#)). The number of patients in 2016-, 2017-, 2018-, and 2019-cohorts is 29,301, 39,553, 50,618, and 63,541, respectively. However, it is important to note that the dataset is highly imbalanced, as indicated by the ratios of decedents to survivors in each cohort: 349:28,952 for 2016, 551:39,000 for 2017, 727:49,891 for 2018, and 783:62,758 for

Table 2. AUC estimates (mean \pm standard deviation) for each classifier, calculated over 5-fold cross-validation applied to the yearly-specific training sets

Classifier	AUC			
	2016	2017	2018	2019
LRR	0.779 \pm 0.014	0.790 \pm 0.019	0.772 \pm 0.014	0.793 \pm 0.010
PER	0.613 \pm 0.070	0.614 \pm 0.069	0.606 \pm 0.079	0.657 \pm 0.071
GNB	0.766 \pm 0.023	0.778 \pm 0.011	0.754 \pm 0.011	0.775 \pm 0.01
SVM	0.780 \pm 0.015	0.789 \pm 0.013	0.772 \pm 0.013	0.791 \pm 0.010
KNN	0.558 \pm 0.018	0.568 \pm 0.008	0.555 \pm 0.012	0.573 \pm 0.013
RF	0.646 \pm 0.031	0.630 \pm 0.015	0.620 \pm 0.018	0.635 \pm 0.018
XGB	0.771 \pm 0.022	0.774 \pm 0.016	0.772 \pm 0.014	0.782 \pm 0.017
LGB	0.775 \pm 0.018	0.783 \pm 0.019	0.767 \pm 0.016	0.788 \pm 0.014
GBRT	0.776 \pm 0.021	0.788 \pm 0.017	0.777 \pm 0.014	0.789 \pm 0.019
ADB	0.785 \pm 0.017	0.788 \pm 0.019	0.772 \pm 0.012	0.791 \pm 0.011
LDA	0.780 \pm 0.016	0.789 \pm 0.014	0.771 \pm 0.014	0.791 \pm 0.018
QDA	0.775 \pm 0.013	0.788 \pm 0.013	0.772 \pm 0.015	0.791 \pm 0.010

2019. **Table 1** contains clinical variables used for predicting mortality. We handled missing values by imputing the median for numeric data and mode for categorical data using the values from the training data. Furthermore, stratified random sampling was used to divide each-year-specific cohort into a training set (80% of the data) and a test set (20% of the data)–this practice maintains the proportion of alive versus deceased cases the same as in the full cohort. The training set is utilized to train and select the predictive model, which is then evaluated using the test set.

Yearly-Specific Classifier Training and Selection for One-Year Mortality Prediction

Twelve different classifiers were utilized in this study: linear models including logistic regression with L_2 ridge penalty (LRR) [17], support vector machines with linear kernel (SVM) [18], linear discriminant analysis (LDA) [19], and perceptron (PER) [20]; Gaussian naïve Bayes (GNB) [20], ensemble methods including RF [21], XGB [22], LightGBM (LGB) [23], gradient boosting with regression trees (GBRT) [24], and Adaboost with decision trees (ADB) [25]; KNN [20]; quadratic discriminant analysis (QDA) [19]. The selection of these classifiers is discussed in detail in the Materials and Methods section. The training process involved selecting the model and tuning its hyperparameters on each yearly-specific training set using stratified 5-fold cross-validation (5-fold CV). The area under the curve (AUC) was chosen as a performance indicator for 5-fold CV. Further information on the search space of hyperparameters for each classifier can be found in the Materials and Methods section. **Table 2** provides the mean and standard deviation of the AUC estimates obtained during 5-fold CV for each classifier. The results indicate that the ADB, LRR, GBRT, and LRR classifiers demonstrated the highest AUC values for the 2016-, 2017, 2018-, and 2019-cohorts, respectively.

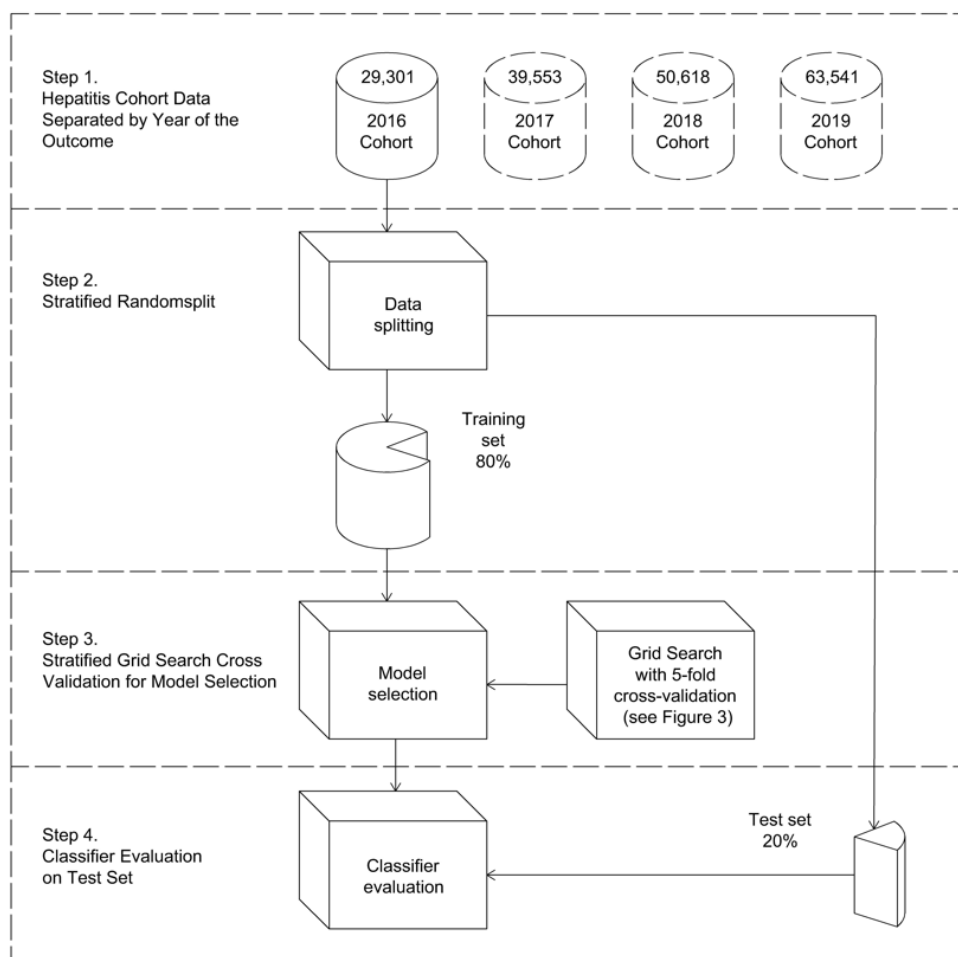


Figure 1. A flow chart describing the constructed machine learning platform (Source: Authors' own elaboration)

Table 3. Performance evaluation of the optimal yearly-specific classifier estimated on their corresponding test sets

Year	Optimal classifier	Balanced accuracy	AUC	Specificity	Sensitivity	G-mean
2016	AdaBoost	0.726	0.793	0.666	0.786	0.723
2017	Logistic regression	0.690	0.771	0.689	0.691	0.691
2018	Gradient boosting with regression trees	0.685	0.746	0.728	0.641	0.684
2019	Logistic regression	0.695	0.787	0.678	0.713	0.695

Evaluation of Year-Specific Classifiers for One-Year Mortality Prediction

The final classifier for each sub-cohort was trained using the optimal year-specific classifier and its hyperparameters identified during the model selection phase. We evaluated selected classifiers based on the respective yearly-specific test sets, measuring performance through metrics such as balanced accuracy, AUC, specificity, sensitivity, and geometric mean score (G-mean).

Figure 1 illustrates the step-by-step process of selecting the best predictive model and evaluating its performance. **Table 3** provides the performance results obtained on the held-out test data. Additionally, **Table A1**, **Table A2**, **Table A3**, and **Table A4** in **Appendix A** include the confusion matrices for each year-specific classifier, based on their test set evaluations. Each classifier reached an AUC over 0.74, which is considered "fair" based on the objective metrics of diagnostic tests. Notably, classifiers from the years 2016 and 2019 obtained an AUC over 0.78, approaching the "good" performance level (as defined in [26]). Furthermore, the results indicate that each classifier, except for the 2018-specific classifier, exhibited greater sensitivity compared to specificity. For our specific

application, high sensitivity is a desirable feature as the risk of failing to identify who is at risk of dying within a year is more critical than mislabeling a patient as at risk of "death" who is likely to survive.

Impact Direction and Importance of Each Feature for One-Year Mortality Prediction

Our approach included conducting a SHapley Additive exPlanations (SHAP) [27] analysis to achieve two main objectives: firstly, to assess the individual significance of each feature in predicting mortality; and secondly, to understand how each feature influences the direction of the prediction. SHAP values were computed for each year-specific classifier selected during the model selection phase. In particular, we computed SHAP values for the ADB classifier in the 2016-cohort, the LRR classifier in the 2017- and 2019-cohorts, and the GBRT classifier in the 2018-cohort. It is essential to note that since no feature selection was conducted, SHAP values were calculated for all clinical variables in the training dataset (see discussion section for further details). The bee swarm plot of SHAP values and bar plot of mean absolute SHAP values for the 2018 year-specific cohort are depicted in left part and right part in **Figure 2**, respectively. SHAP plots for the other year-

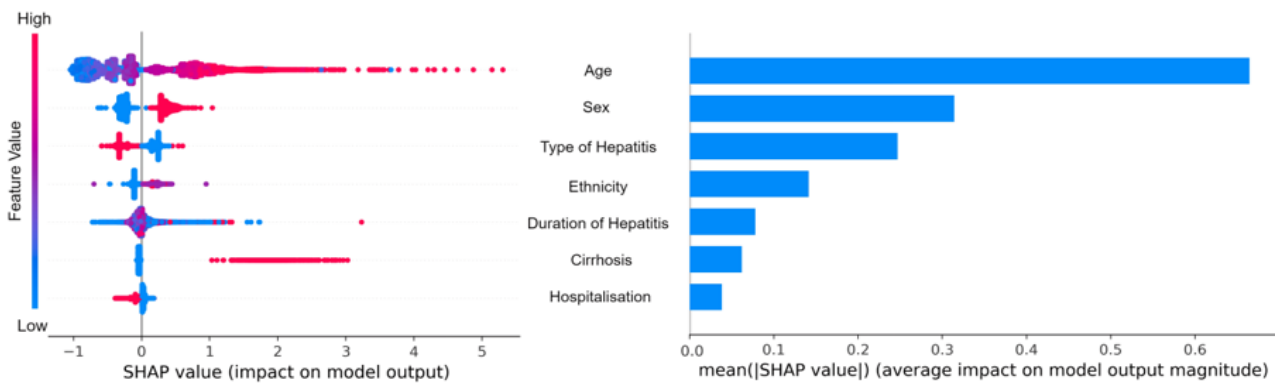


Figure 2. SHAP results for a cohort of 2018 (left) bee swarm plot of SHAP values for the 2018 year cohort (red points show high feature values for a patient, while blue points denote low feature values & red points with positive SHAP values show direct relationship between the feature and the outcome, while the blue points with the same positive values suggest an inverse relationship) & (right) bar plot of the mean absolute SHAP values for the 2018 year cohort (bar plot illustrates the importance of each feature in predicting the outcome, where longer bars indicate a higher significance) (Source: Authors' own elaboration)

specific cohorts can be found in **Appendix A**. From the right part in **Figure 2** it can be observed that age is the most important predictive feature and the left part in **Figure 2** shows that older age has an association with higher mortality. To get a summary of the SHAP values across cohorts, we computed the *average* of the mean absolute SHAP values (AMAS) for each feature. By computing the AMAS value for each feature, we have obtained the following ranking of the features (in the order of importance): age, sex, type of hepatitis, ethnicity, duration of hepatitis, cirrhosis, and hospitalization. The corresponding AMAS values for these features were 0.697, 0.298, 0.121, 0.099, 0.084, 0.039, and 0.020, respectively. Obtained findings indicate four key factors—age, sex, type of hepatitis, and ethnicity—rank as the most critical features for one-year mortality prediction.

DISCUSSION

Several studies have utilized a combination of administrative data (such as demographics and comorbidities) and clinical data (including vital signs, laboratory results) for mortality prediction of CVH patients. For instance, it was used the NB, C4.5 classifier, and decision table to assess the risks of hepatitis disease [28]. Among these algorithms, the NB classifier demonstrated the best performance, achieving an f-measure of 0.848, and a sensitivity of 0.853. It was utilized NB, DT, SVM, and LR to predict HCV patient mortality. The LR algorithm outperformed the other models achieving an f-measure of 0.86 and a sensitivity of 0.87 [29]. Another study focused on predicting the mortality of hepatitis patients using the LR algorithm, which showed an f-measure of 0.75 and a sensitivity of 0.9 [30]. It was predicted the mortality in patients with HBV using six classifiers including DT, LR, SVM, RF, ADB, and XGB [13]. Both ADB and LR outperformed the other classifiers, achieving an AUC of 0.93. Notably, among the reviewed studies, only it was reported an AUC and considered model explainability by performing SHAP analysis [13]. Results from their study indicated that bilirubin, high ascites levels, age, alkaline, and malaise levels are important features, with bilirubin being the most significant one.

In comparison to these studies, our predictive models achieved lower AUC estimates. In particular, **Table 3** demonstrates that classifiers developed for each specific year reached an AUC in the range of 0.74 to 0.8—according to [26], an

AUC in the range of 0.7 to 0.8 is considered ‘fair’ for a diagnostic test. Lower AUC estimates in our study can be partly explained due to the “administrative” nature of our features, which are generally easy to collect. In particular, in contrast with these studies, we neither use vital indicators nor laboratory tests.

The findings of our research demonstrate that the four key predictors of one-year mortality for patients with CVH are age, gender, hepatitis type, and ethnicity. The left part in **Figure 2**, as well as **Figure A1**, **Figure A2**, and **Figure A3** in **Appendix A**, illustrate a noticeable direct relationship between higher mortality and predictors such as older age and male sex. Furthermore, from the left part in **Figure 2**, it is observed that HCV patients have a higher risk of mortality than HBV patients. These findings are consistent with previous studies [13, 31, 32].

In particular, the association between older age and higher mortality in CVH patients has been established by several studies [31, 33]. In another work, it was examined the relationship between age groups (20-49, 50-64, 65-85 years) and mortality of patients with HCV based on data collected from the veterans health administration hepatitis C clinical case registry of the United States [34]. The findings showed that patients in the older age groups (50-64 and 65-85 years) have higher mortality rates compared to patients in the younger age group (20-49 years). Similarly, research conducted on a nationwide register-based cohort in Denmark demonstrated that patients with HBV at an older age exhibit an elevated risk of mortality in comparison to younger patients [32].

Several studies have examined the association between sex and hepatitis mortality. A population-based cohort study from France found that male patients with HBV have higher all-cause and HBV-related mortality rates than their female counterparts [35]. The outcome of our research aligns with this study, demonstrating that males with CVH face an elevated risk of mortality compared to females diagnosed with the same disease. That being said, some studies have observed a greater risk of all-cause mortality in female patients with HCV [36].

A comprehensive 13-year population-based study in Asia found that patients with HCV faced a substantially increased risk of cardiovascular outcomes and overall mortality compared to those diagnosed with HBV [37]. Our study also showed a similar tendency, supporting these findings. Another factor of interest is the association between ethnicity and mortality among patients. Several studies investigated the prevalence and mortality of hepatitis in developed countries,

considering ethnic disparities [38-40]. For example, a study conducted based on the chronic hepatitis cohort study from the United States revealed that African Americans had the highest rates of mortality, 26% higher than white patients, whereas Asian American/American Indian/Pacific Islander patients had the lowest mortality rates [40]. Similar to our conclusion, previous research based on the UNEHS database also reported a noticeable difference of mortality rates depending on ethnicity groups [41, 42].

Although our investigation utilized a relatively limited number of administrative features (seven variables detailed in **Table 1**), the constructed classifiers achieved an AUC in the ‘fair’ range, which is a noticeable accomplishment in predicting one-year mortality in CVH patients. Another strength of our work is that the data is collected from a population-based registry and for a sufficiently long time that enables constructing classifiers based on a true random sample of the population. Moreover, this research is the first of its kind in Central Asia, providing valuable insights on predicting one-year mortality rates of hepatitis patients in the region. The findings and the constructed models can help in creating better treatment plans and approaches for managing hepatitis in various healthcare environments. The findings could also be useful in raising public awareness and promoting healthier lifestyles to reduce hepatitis-related mortality.

Our study has several limitations. From a clinical perspective, our study does not include laboratory data and detailed medical histories of the patients. Moreover, important information regarding comorbidities, such as diabetes mellitus, HIV, cardiovascular diseases, renal diseases, and the precise timestamp attached to each comorbidity, as well as anthropometric indices (BMI), and alcohol use, were not considered. Including these details could enhance the accuracy and effectiveness of predictive models developed in the future. However, incorporating such data would incur additional costs.

From the standpoint of ML, our study lacks a feature selection stage. Although, this stage is less crucial in our current study due to the limited number of features and the large sample size, the inclusion of clinical notes or laboratory data could introduce additional features. In such a scenario, it would be generally expected to have a feature selection stage to mitigate the challenges related to high-dimensional data (also referred to as the peaking phenomenon [43] in pattern recognition). These aspects will be the subject of our future investigations.

CONCLUSION

In this study, an advanced ML platform was developed to predict one-year mortality in CVH patients using data from administrative health records. The constructed classifiers achieved an AUC in the range from 0.74 to 0.8, rated as ‘fair’ and approaching ‘good’, according to standard diagnostic test metrics. The AUC results indicate the feasibility of using solely low-cost administrative health data to predict one-year mortality in CVH patients. Moving forward, combining this data with key comorbidities, laboratory data, body measurements, and patients’ medical history could potentially lead to more accurate and robust predictive models, further enhancing patient care and treatment outcomes. The study identified that the top four most important predictors are age, sex, type of

hepatitis, and ethnicity. These findings have significant implications, potentially leading to better tailored treatment approaches for hepatitis patients and could also support public health initiatives and encourage the adoption of healthier lifestyles to prevent hepatitis-related mortality.

MATERIALS AND METHODS

Study Population

The initial dataset obtained from UNEHS included a substantial collection of 20,810,911 medical records from UNEHS, covering years 2014 to 2019, as two separate registries—inpatient and outpatient. Among the 11,157,509 records in the outpatient registry, 69,560 were unique patients with CVH, identified using the international classification of diseases 10th revision (ICD-10) codes for hepatitis, specifically B18.1 (CVH B without delta-agent) and B18.2 (CVH C). The inpatient registry comprised a total of 9,653,402 records, involving 20,170 unique hepatitis patients (see **Appendix A** for details on how patients were selected). After combining the two registries and removing duplicates, the final cohort consisted of 82,700 unique hepatitis patients. Ethical approval was obtained from the Nazarbayev University Institutional Review Ethics Committee (NU-IREC) #745/12062023. As the study was conducted with secondary data from UNEHS, no informed consent was obtained. All research methods followed the “Reporting of studies conducted using observational routinely collected health data” (RECORD) guideline.

Data Preprocessing

Patients with missing vital outcomes were excluded from our analysis. The remaining data was organized into four year-specific sub-cohorts: for the years 2016, 2017, 2018, and 2019. Each sub-cohort was used to predict mortality for the corresponding year, using the clinical data gathered until the end of the previous year. The number of patients in each cohort was, as follows: 29,301 in 2016-cohort, 39,553 in 2017-cohort, 50,618 in 2018-cohort and 63,541 in 2019-cohort, respectively. However, it is important to note that the dataset is highly imbalanced, as indicated by the ratios of decedents to survivors in cohorts: 349:28,952 for 2016, 551:39,000 for 2017, 727:49,891 for 2018, and 783:62,758 for 2019. There were seven clinical variables, including age, sex, type of hepatitis, duration of hepatitis, cirrhosis, and hospitalization, to predict mortality. For handling missing data, we imputed numeric feature values using the median of the corresponding variables in the training data, while missing categorical feature values were imputed using the mode. Finally, each year-specific cohort was randomly divided into training and test sets in an 80/20 ratio using stratified sampling to ensure the same proportions of decedents and survivors as in the complete cohort.

Model Training, Selection, and Evaluation

Twelve different classifiers were utilized in this study: linear models including logistic regression with L_2 ridge penalty (LRR), support vector machines with linear kernel (SVM), LDA, and PER; GNB; ensemble methods including RF, XGB, LightGBM (LGB), gradient boosting with regression trees (GBRT), and Adaboost with decision trees (ADB); KNN; quadratic discriminant analysis (QDA). **Table 4** displays the hyperparameter values that were utilized during the model selection phase for these classifiers.

Table 4. Search space of hyperparameters for model selection using grid search with cross-validation

Classifiers	Hyperparameter	Candidate hyperparameter space
LRR	Penalty	L_2
	Regularization parameter C	100, 10, 1.0, 0.1, 0.01
PER	Alpha	0.0001, 0.001, 0.01
	Penalty	L_2, L_1, none
GNB	-	-
RF	Number of estimators	10, 100, 1000
	Maximum depth	2, 5, 10, 20, 50
	Maximum features	'auto', 'sqrt', 'log2'
	Maximum depth	5, 10, 100
XGB	Number of estimators	10, 100, 1000
	Learning rate	0.001, 0.01, 0.1
	Maximum depth	5, 10, 100
LGB	Number of estimators	10, 100, 1000
	Learning rate	0.001, 0.01, 0.1
GRBT	Number of estimators	10, 100, 1000
	Learning rate	0.001, 0.01, 0.1
AdaBoost	Number of estimators	10, 100, 1000
	Learning rate	0.001, 0.01, 0.1
KNN	Number of neighbours	3, 5
SVM	Penalty	L_2
	Kernel	Linear
	Regularization parameter C	0.1, 0.5, 1, 5
LDA	Solver	'svd', 'lsqr', 'eigen'
	Tolerance	0.00001, 0.0001, 0.0003
Quadratic discriminant analysis	Regularization parameter	0.1, 0.5, 0.7, 0.9
	Tolerance	0.00001, 0.0001, 0.0003

The selection of predictive models was influenced by previous studies, in which these models had been commonly employed to predict hepatitis infection, treatment response, and mortality rate. It was predicted the infected patients with HBV using NB, KNN, RF, and LR [7]. In another work, DT, LR, SVM, RF, ADB, and XGB were used to predict mortality rate of HBV patients [13]. In another study, RF, SVM, and LR were utilized to predict 30-day and 90-day mortality rates of patients diagnosed with alcoholic hepatitis [44]. LGB and XGB were employed for pre-diagnosis of acute liver failure in [45]. XGB is regarded as one of the leading models for processing tabular data and has been extensively utilized for Kaggle competitions [46].

We employ stratified 5-fold cross-validation (5-fold CV) to select the best predictive model on each yearly-specific training set. The entire procedure of model selection using a 5-fold CV is illustrated in **Figure 3**. The AUC is used as the performance metric for selecting a year-specific classifier, as the AUC is not reliant on any particular decision threshold in the classifier. Moreover, the decision threshold is further tuned to achieve the highest G-mean, especially in highly imbalanced datasets where a 'default' decision threshold could result in low G-mean scores. The fine-tuning process was done by varying the threshold between 0 and 1 in increments of 0.001 and calculating the G-mean at each threshold point.

The optimal year-specific classification rule, along with its hyperparameter values determined through 5-fold cross-validation, was utilized to train a final year-specific classifier on the complete normalized training set.

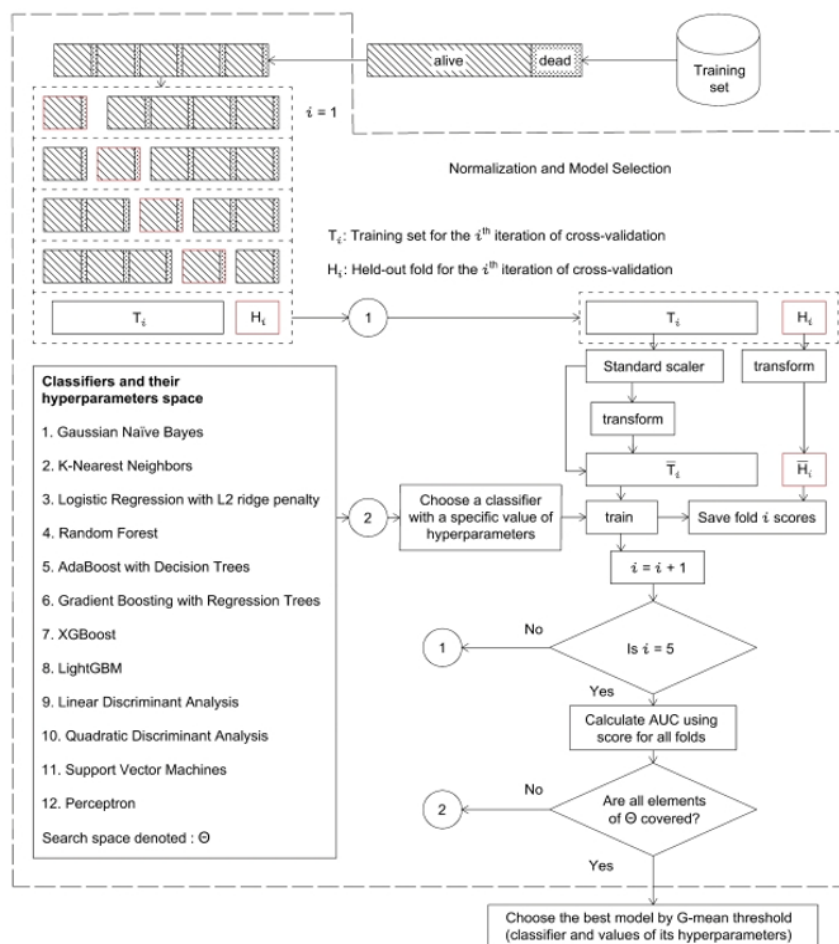


Figure 3. Diagram illustrating the model selection implemented using 5-fold cross-validation (Source: Authors' own elaboration)

Software and Packages

The computational work for this study was conducted on a virtual server, which was powered by an AMD Opteron Processor 6174 at 2.19 GHz. This server was equipped with 22 processors, a total storage capacity of 3.9 TB, and 200 GB of RAM. The main program was done in Python (version 3.11; Python Software Foundation), utilizing open-source packages such as scikit-learn, pandas, seaborn, matplotlib, XGB, lightgbm, and shap.

Author contributions: **IA:** implemented ML framework and contributed to drafting the manuscript; **DS:** involved in data management, offered clinical expertise, and assisted in manuscript preparation; **AA:** assisted in implementing the ML framework; **GZ, YS, & SY:** participated in managing the data; **AZ:** designed the ML framework, drafting the manuscript, and took part in the experimental design and coordination; **AG:** initiated the study, provided clinical insights, and played a key role in both coordinating the study and drafting the manuscript. All authors have agreed with the results and conclusions.

Funding: This study was supported by grants from the Nazarbayev University Faculty Development Competitive Research Grant Program (AI and Data Science) 2024-2026 (Funder Project Reference: 201223FD2604, title: An AI-based approach for analyzing electronic medical records: Prediction of healthcare outcomes and drug demand). A.G. is a PI of the project.

Acknowledgments: The authors would like to thank all staff from the Republican Center of Electronic Healthcare for providing data and consultancy.

Ethical statement: The authors stated that the study was approved by the Nazarbayev University Institutional Review Ethics Committee (NU-IREC #745/12062023). Written informed consents were obtained from the participants.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

- Lai CL, Ratziu V, Yuen MF, Poynard T. Viral hepatitis B. *Lancet*. 2003;362(9401):2089-94. [https://doi.org/10.1016/S0140-6736\(03\)15108-2](https://doi.org/10.1016/S0140-6736(03)15108-2) PMID:14697813
- Poynard T, Yuen MF, Ratziu V, Lai CL. Viral hepatitis C. *Lancet*. 2003;362(9401):2095-100. [https://doi.org/10.1016/S0140-6736\(03\)15109-4](https://doi.org/10.1016/S0140-6736(03)15109-4) PMID:14697814
- WHO. Global progress report on HIV, viral hepatitis and sexually transmitted infections. World Health Organization; 2021. Available at: <https://www.who.int/publications/item/9789240027077> (Accessed: 10 June 2023)
- Ashimkhanova A, Syssoyev D, Gusmanov A, et al. Epidemiological characteristics of chronic viral hepatitis in Kazakhstan: Data from unified nationwide electronic healthcare system 2014-2019. *Infect Drug Resist*. 2022;15:3333-46. <https://doi.org/10.2147/IDR.S363609> PMID:35782528 PMCid:PMC9248955
- WHO. Combating hepatitis B and C to reach elimination by 2030. World Health Organization; 2021. Available at: <https://apps.who.int/iris/handle/10665/206453> (Accessed: 10 June 2023)
- Li THS, Chiu HJ, Kuo PH. Hepatitis C virus detection model by using random forest, logistic regression, and ABC algorithm. *IEEE Access*. 2022;10:91045-58. <https://doi.org/10.1109/ACCESS.2022.3202295>
- Mamdouh Farghaly H, Shams MY, Abd El-Hafeez T. Hepatitis C virus prediction based on machine learning framework: A real-world case study in Egypt. *Knowl Inf Syst*. 2023;65:2595-617. <https://doi.org/10.1007/s10115-023-01851-4>
- Alizargar A, Chang YL, Tan TH. Performance comparison of machine learning approaches on hepatitis C prediction employing data mining techniques. *Bioengineering (Basel)*. 2023;10(4):481. <https://doi.org/10.3390/bioengineering10040481> PMID:37106668 PMCid:PMC10135598
- Haga H, Sato H, Koseki A, et al. A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus. *PLoS One*. 2020;15(11):e0242028. <https://doi.org/10.1371/journal.pone.0242028> PMID:33152046 PMCid:PMC7644079
- Kashif AA, Bakhtawar B, Akhtar A, et al. Treatment response prediction in hepatitis C patients using machine learning techniques. *Int J Technol Innov Manag*. 2021;1(2):79-89. <https://doi.org/10.54489/ijtim.v1i2.24>
- Tian X, Chong Y, Huang Y, et al. Using machine learning algorithms to predict hepatitis B surface antigen seroclearance. *Comput Math Methods Med*. 2019;2019:6915850. <https://doi.org/10.1155/2019/6915850> PMID:31281411 PMCid:PMC6594274
- Butt MB, Alfayad M, Saqib S, et al. Diagnosing the stage of hepatitis C using machine learning. *J Healthc Eng*. 2021;2021:8062410. <https://doi.org/10.1155/2021/8062410> PMID:35028114 PMCid:PMC8748759
- Obaido G, Ogbuokiri B, Swart TG, et al. An interpretable machine learning approach for hepatitis B diagnosis. *Appl Sci*. 2022;12(21):11127. <https://doi.org/10.3390/app122111127>
- Albogamy FR, Asghar J, Subhan F, et al. Decision support system for predicting survivability of hepatitis patients. *Front Public Health*. 2022;10:862497. <https://doi.org/10.3389/fpubh.2022.862497> PMID:35493354 PMCid:PMC9051027
- Ali N, Srivastava D, Tiwari A, Pandey AK, Sahu A. Predicting life expectancy of hepatitis B patients using machine learning. In: Proceedings of the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics. 2022. p. 1-4. <https://doi.org/10.1109/ICDCECE53908.2022.9793025>
- Gusmanov A, Zhakhina G, Yerdessov S, et al. Review of the research databases on population-based registries of unified electronic healthcare system of Kazakhstan (UNEHS): Possibilities and limitations for epidemiological research and real-world evidence. *Int J Med Inform*. 2023;170:104950. <https://doi.org/10.1016/j.ijmedinf.2022.104950> PMID:36508752
- Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. London: Springer; 2009. <https://doi.org/10.1007/978-0-387-84858-7>
- Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-97. <https://doi.org/10.1007/BF00994018>
- Anderson TW. Classification by multivariate analysis. *Psychometrika*. 1951;16(1):31-50. <https://doi.org/10.1007/BF02313425>
- Duda RO, Hart PE, Stork DG. Pattern classification. Hoboken: John Wiley & Sons; 2001.
- Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. <https://doi.org/10.1023/A:1010933404324>

22. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785-94. <https://doi.org/10.1145/2939672.2939785>
23. Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-54.
24. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-232. <https://doi.org/10.1214/aos/1013203451>
25. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55(1):119-39. <https://doi.org/10.1006/jcss.1997.1504>
26. Pines JM, Carpenter CR, Raja AS, Schuur JD. Evidence-based emergency care: Diagnostic testing and clinical decision rules. Hoboken: John Wiley & Sons; 2012. <https://doi.org/10.1002/9781118482117>
27. Lundberg SM, Allen PG, Lee SI. A unified approach to interpreting model predictions. In: *Advances in neural information processing systems*. Newry: Curran Associates Inc; 2017.
28. Yildirim P. Filter-based feature selection methods for prediction of risks in hepatitis disease. *Int J Mach Learn Comput*. 2015;5(4):258-63. <https://doi.org/10.7763/IJMLC.2015.V5.517>
29. Bhargav KS, Thota D, Kumari TD, Vikas B. Application of machine learning classification algorithms on hepatitis dataset. *Int J Appl Eng Res*. 2018;13(16):12732-7.
30. Nivaan GV, Emanuel AWR. Analytic predictive of hepatitis using the regression logic algorithm. In: *Proceedings of the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems*. 2020. p. 106-10. <https://doi.org/10.1109/ISRITI51436.2020.9315365>
31. Fedeli U, Grande E, Grippo F, Frova L. Mortality associated with hepatitis C and hepatitis B virus infection: A nationwide study on multiple causes of death data. *World J Gastroenterol*. 2017;23(10):1866-76. <https://doi.org/10.3748/wjg.v23.i10.1866> PMID:28348493 PMCID:PMC5352928
32. Bollerup S, Hallager S, Engsig F, et al. Mortality and cause of death in persons with chronic hepatitis B virus infection versus healthy persons from the general population in Denmark. *J Viral Hepat*. 2022;29(8):727-36. <https://doi.org/10.1111/jvh.13713> PMID:35633092
33. Alavi M, Grebely J, Hajarizadeh B, et al. Mortality trends among people with hepatitis B and C: A population-based linkage study, 1993-2012. *BMC Infect Dis*. 2018;18(1):215. <https://doi.org/10.1186/s12879-018-3110-0> PMID:29743015 PMCID:PMC5944091
34. El-Serag HB, Kramer J, Duan Z, Kanwal F. Epidemiology and outcomes of hepatitis C infection in elderly US Veterans. *J Viral Hepat*. 2016;23(9):687-96. <https://doi.org/10.1111/jvh.12533> PMID:27040447
35. Montuclard C, Hamza S, Rollot F, et al. Causes of death in people with chronic HBV infection: A population-based cohort study. *J Hepatol*. 2015;62(6):1265-71. <https://doi.org/10.1016/j.jhep.2015.01.020> PMID:25625233
36. Ireland G, Mandal S, Hickman M, Ramsay M, Harris R, Simmons R. Mortality rates among individuals diagnosed with hepatitis C virus (HCV): An observational cohort study, England, 2008 to 2016. *Euro Surveill*. 2019;24(30):1800695. <https://doi.org/10.2807/1560-7917.ES.2019.24.30.1800695> PMID:31362807 PMCID:PMC6668288
37. Wu VC-C, Chen T-H, Wu M, et al. Comparison of cardiovascular outcomes and all-cause mortality in patients with chronic hepatitis B and C: A 13-year nationwide population-based study in Asia. *Atherosclerosis*. 2018;269:178-84. <https://doi.org/10.1016/j.atherosclerosis.2018.01.007> PMID:29366991
38. Emmanuel B, Shardell MD, Tracy L, Kottitil S, El-Kamary SS. Racial disparity in all-cause mortality among hepatitis C virus-infected individuals in a general US population, NHANES III. *J Viral Hepat*. 2017;24(4):380-8. <https://doi.org/10.1111/jvh.12656> PMID:27905175 PMCID:PMC5739320
39. Bixler D, Zhong Y, Ly KN, et al. Mortality among patients with chronic hepatitis B infection: The chronic hepatitis cohort study (CHeCS). *Clin Infect Dis*. 2019;68(6):956-63. <https://doi.org/10.1093/cid/ciy598> PMID:30060032 PMCID:PMC11230463
40. Lu M, Li J, Zhou Y, et al. Trends in cirrhosis and mortality by age, sex, race, and antiviral treatment status among US chronic hepatitis B patients (2006-2016). *J Clin Gastroenterol*. 2022;56(3):273-9. <https://doi.org/10.1097/MCG.0000000000001522> PMCID:PMC10257940
41. Yerdessov S, Almukhambetova A, Mambetaliev M, et al. Epidemiological characteristics and climatic variability of viral meningitis in Kazakhstan, 2014-2019. *Front Public Health*. 2023;10:1041135. <https://doi.org/10.3389/fpubh.2022.1041135> PMID:36684964 PMCID:PMC9845948
42. Midlenko A, Mussina K, Zhakhina G, et al. Prevalence, incidence, and mortality rates of breast cancer in Kazakhstan: Data from the Unified National Electronic Health System, 2014-2019. *Front Public Health*. 2023;11:1132742. <https://doi.org/10.3389/fpubh.2023.1132742> PMID:37143985 PMCID:PMC10153091
43. Zollanvari A, James AP, Sameni R. A theoretical analysis of the peaking phenomenon in classification. *J Classif*. 2020; 37(2):421-34. <https://doi.org/10.1007/s00357-019-09327-3>
44. Gao B, Wu T-C, Lang S, et al. Machine learning applied to omics datasets predicts mortality in patients with alcoholic hepatitis. *Metabolites*. 2022;12(1):41. <https://doi.org/10.3390/metabo12010041> PMID:35050163 PMCID:PMC8781791
45. Zhang D, Gong Y. The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. *IEEE Access*. 2020;8:220990-220003. <https://doi.org/10.1109/ACCESS.2020.3042848>
46. Brownlee J. XGBoost with Python: Gradient boosted trees with XGBoost and scikit-learn. San Fransisco: Machine Learning Mastery; 2018.

APPENDIX A: SUPPLEMENTARY MATERIALS

Supplementary Tables

Table A1. Confusion matrix for 2016-year cohort test set

		Predicted	
		Negative	Positive
Actual	Negative	True negative = 3,857	False positive = 1,934
	Positive	False negative = 15	True positive = 55

Table A2. Confusion matrix for 2017-year cohort test set

		Predicted	
		Negative	Positive
Actual	Negative	True negative = 5,376	False positive = 2,425
	Positive	False negative = 34	True positive = 76

Table A3. Confusion matrix for 2018-year cohort test set

		Predicted	
		Negative	Positive
Actual	Negative	True negative = 7,269	False positive = 2,710
	Positive	False negative = 52	True positive = 93

Table A4. Confusion matrix for 2019-year cohort test set

		Predicted	
		Negative	Positive
Actual	Negative	True negative = 8,504	False positive = 4,048
	Positive	False negative = 45	True positive = 112

Supplementary Figures

SHAP analysis plots

In **Figure A1**, **Figure A2**, and **Figure A3**, a red dot in plots on the left indicates a high value of the feature for a patient, whereas a blue dot represents a low value. Positive SHAP values for red dots show a direct relationship between the feature and the outcome, whereas the same values for blue dots imply an inverse relationship. The direction of SHAP values, positive or negative, corresponds to an increase or decrease in the likelihood of mortality, respectively. The plot on the right illustrates the feature importance on outcome prediction made by the model (a longer bar shows a more significant predictor).

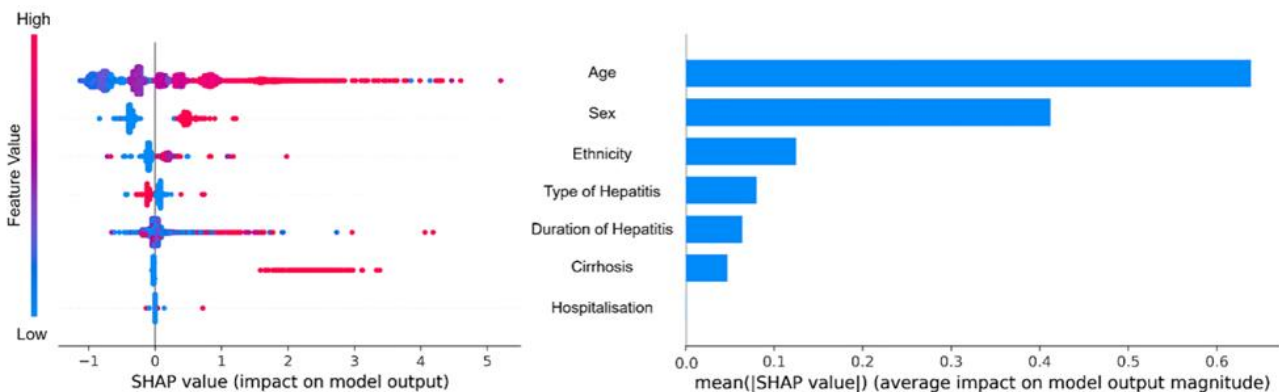


Figure A1. SHAP analysis of 2016-specific cohort: (left) SHAP bee swarm plot & (right) bar plot of the mean absolute SHAP values for 2016-specific cohort (Source: Authors' own elaboration)

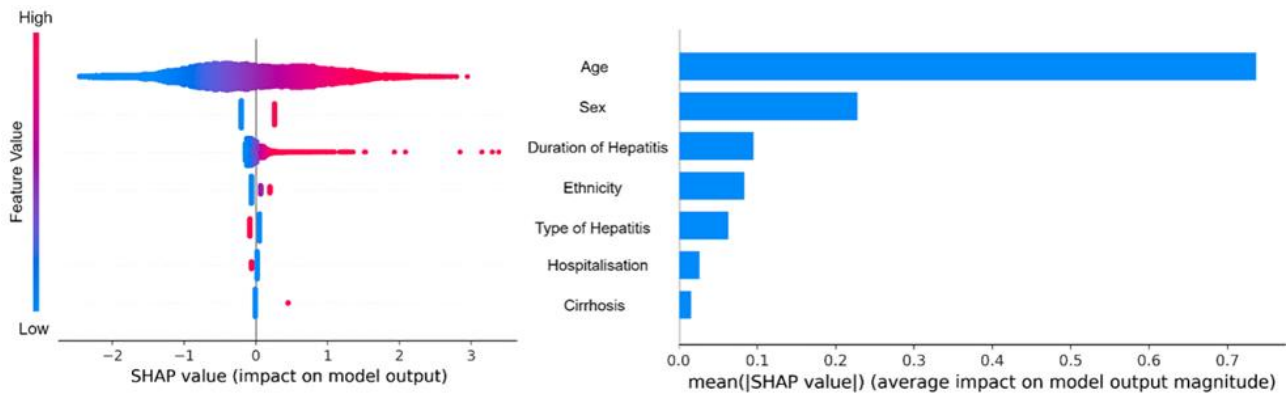


Figure A2. SHAP analysis of 2017-specific cohort: (left) SHAP bee swarm plot & (right) bar plot of the mean absolute SHAP values for 2017-specific cohort (Source: Authors’ own elaboration)

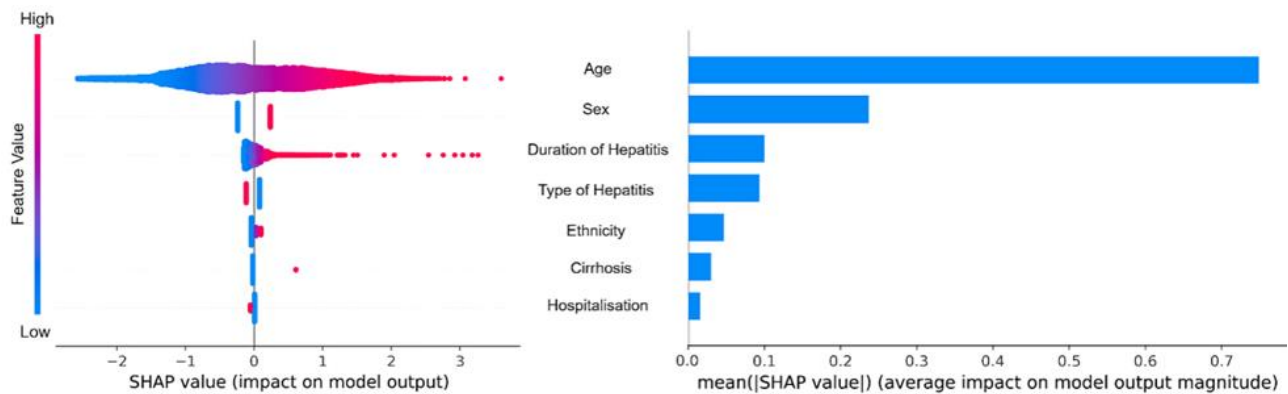


Figure A3. SHAP analysis of 2019-specific cohort: (left) SHAP bee swarm plot & (right) bar plot of the mean absolute SHAP values for 2019-specific cohort (Source: Authors’ own elaboration)

Description of the sub-cohort selection

Figure A4 illustrates six distinct patient groups in our dataset. A triangle indicates the date of hepatitis diagnosis, while a circle indicates the exit date, which signifies the death of a patient. For instance, we choose the year 2018 as the year of observation. The subcohort for 2018 is made out of two patient groups. The first group (case group) includes patients who were diagnosed before the beginning of 2018 and died during that year (similar to case 1 in **Figure A4**, which is identified by a grey line and markers). The second group (control group) comprises patients diagnosed before the beginning of 2018 but were still alive during that year (similar to case 2 and 4, which are also highlighted by grey lines and markers). Patients who died before the start of 2018 were excluded from the subcohort (case 3 in **Figure A4**). Similarly, those who were diagnosed with hepatitis during 2018 were excluded from the subcohort either (case 5 or case 6 in **Figure A4**). Only patients with available clinical information and who were alive up to the end of 2017 were included. Therefore, we chose subcohorts for 2018 and predicted one-year mortality for 50,618 patients out of a total of 82,700 patients. It is noteworthy that hepatitis can occur much earlier than the diagnosis date.

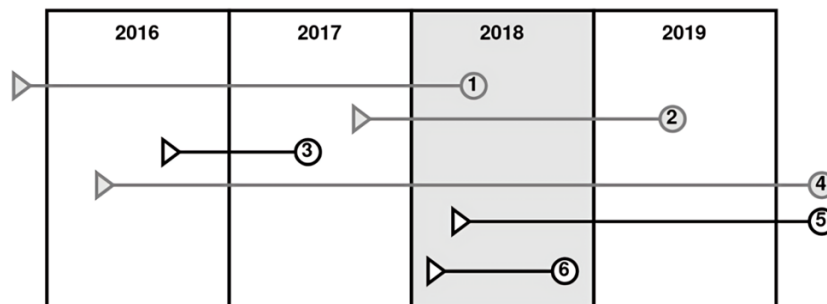


Figure A4. Description of sub-cohort selection (Source: Authors’ own elaboration)

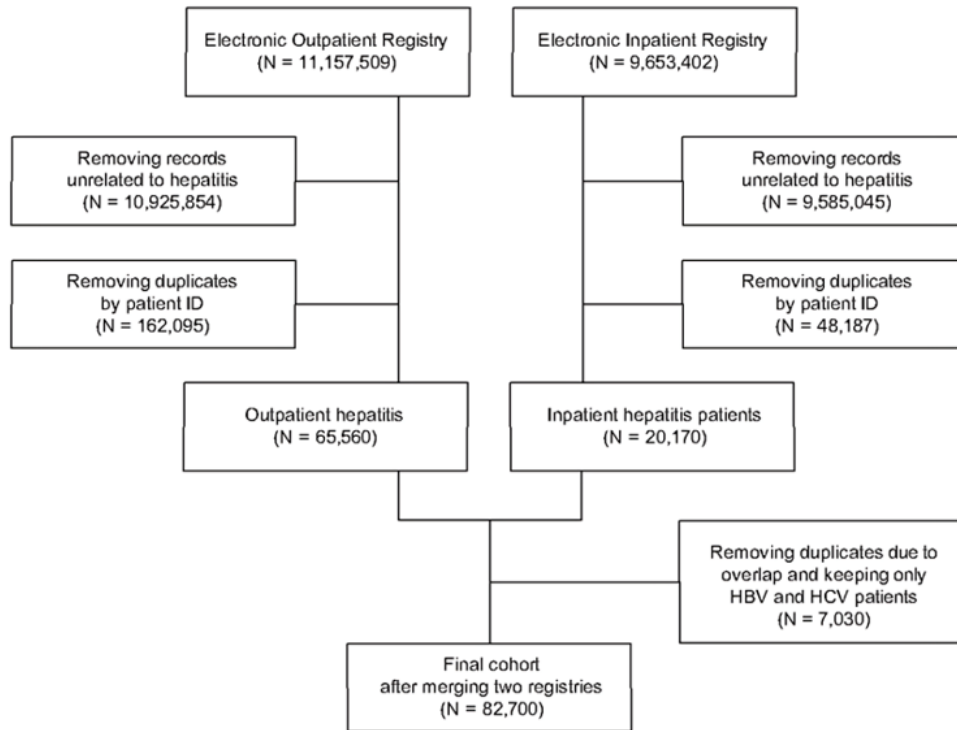


Figure A5. The flowchart of stepwise cohort selection (Source: Authors' own elaboration)