College of Pharmacy
UNIVERSITY OF HOUSTON
Poster # MSR50

Machine Learning Models for Predicting Metabolic Dysfunction-Associated Steatohepatitis (MASH) in the General United States Population

Javeria Khalid, MS[1]; Rajender R. Aparasu, PhD, FAPhA[1]

[1]College of Pharmacy, University of Houston, Texas, United States

Contact Information:
Javeria Khalid
University of Houston
Phone: (346) 473-9476
Email: jkhalid4@uh.edu

BACKGROUND

MASH

- Metabolic dysfunction associated with steatohepatitis (MASH), a severe liver complication, is the second most prevalent reason for liver transplantation in the US [1].
- Current diagnostic methods, relying on invasive procedures and imaging techniques, encounter accessibility challenges.
- Therefore, this study utilized machine learning (ML) techniques to predict MASH using the National Health and Nutrition Examination Survey (NHANES) [2].

Research Gap

- There is no well-performing algorithmic tool for early predicting MASH. [3].
- There are existing studies on the risk factors and prediction risk scores; however, their results are controversial [4].
- Machine learning approaches can be useful in developing the best predictive models for early prediction of MASH [5,6].

OBJECTIVE

The aim of this study was to evaluate machine learning models to predict NASH by using demographic and clinical data on participants diagnosed with NASH by transient elastography.

METHODS

Data sources:

- This retrospective study used the National Health and Nutrition Examination Survey (NHANES) Database from 2017 to 2020.

Inclusion Criteria:

- Age ≥18 years
- Patients for whom valid reproducible Fibro Scan® measurements are available in NHANES.

Exclusion Criteria:

- If participants did not have a FibroScan® controlled attenuation parameter (CAP) or
- If participants were pregnant or were considered high alcohol consumers (average daily consumption of ≥20 g/day and ≥30 g/day for women and men, respectively
- If participants had other potential causes of liver disease, including viral hepatitis (defined as positive for hepatitis A, B or C, D or E) or human immunodeficiency virus (HIV) (reported or serology)

Operational Definition of Target Population:

- NASH was defined based on transient liver ultrasonography using CAP.
- Based on published MASH criterion [10], participants were classified into two groups, i.e.,
  - CAP < 270 dB/m patients without NASH
  - CAP ≥ 270 dB/m is ≥ stage 2 of Steatohepatitis.

Features/Variables included in the analysis:

- All possible risk factors were identified using a literature review.
- Approximately 41 variables available in the NHANES database were included in the model for final analysis.

Analyses:

- Data was initially cleaned in SAS software, and a cohort was created by applying all inclusion and exclusion criteria. Final analysis was done using

Machine Learning algorithm

- Data was divided into 75: 25 ratios (training data = 75% & test data = 25%).
- Keeping in view the binary outcome, the following five machine learning approaches were considered for predictive modeling for MASH patients.
  1. Logistic Regression
  2. K-Nearest Neighbor (KNN)
  3. Super Vector Machine Classification (SVM Classification)
  4. Decision Tree
  5. Random Forest

RESULTS

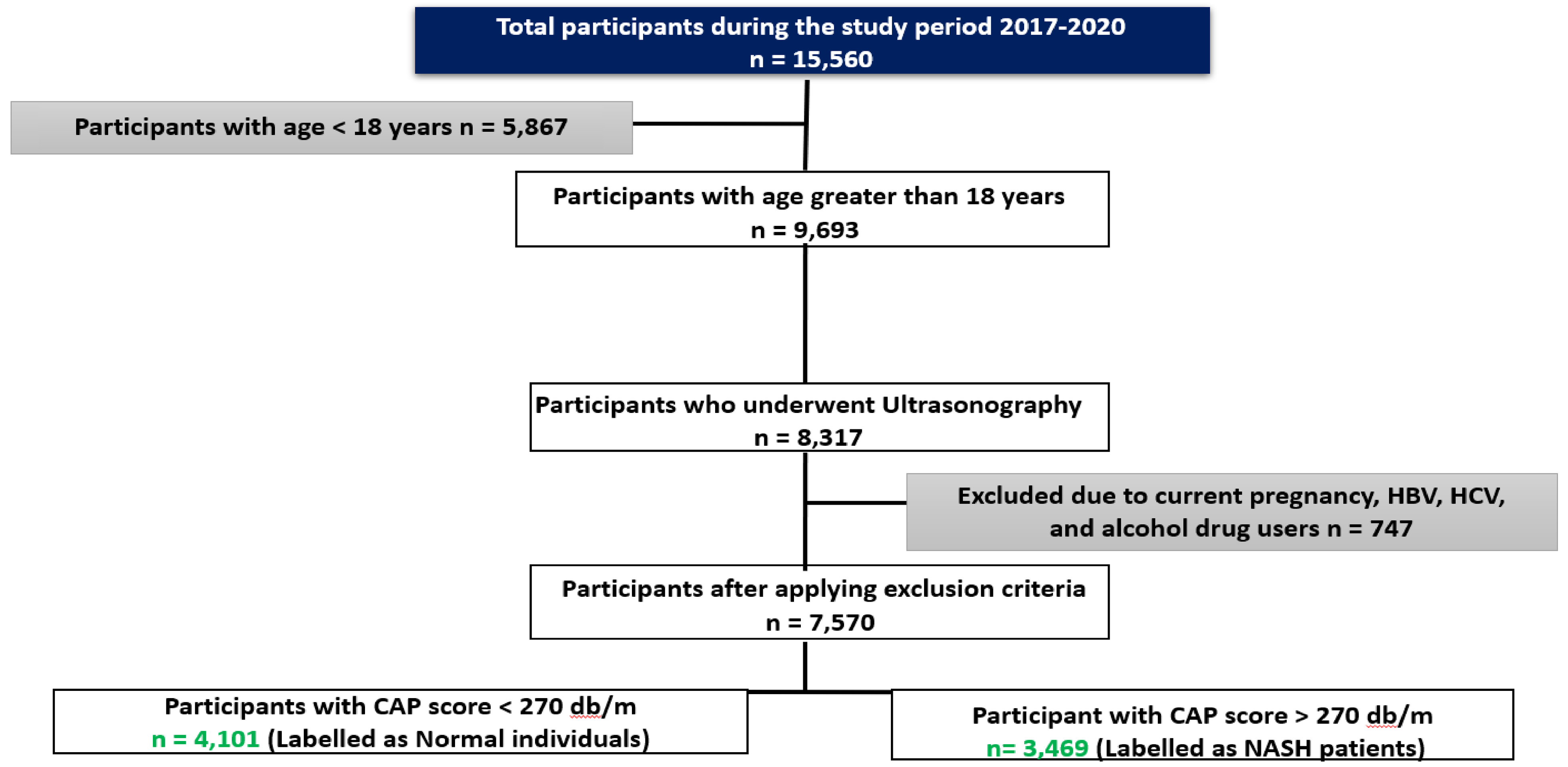

Table 2. Predictive performance of the five machine-learning models

Machine Learning Model	Accuracy	AUROC	Precision	Recall	F1_Score
Logistic	0.752	0.835	0.72	0.73	0.72
KNN	0.7245	0.734	0.69	0.68	0.68
SVM	0.7509	0.84	0.74	0.71	0.73
Decision Tree Classification	0.74	0.69	0.69	0.75	0.72
Random Forest	0.752	0.841	0.72	0.77	0.74

CONCLUSIONS

- The study found that model performance for ML models was similar to the logistic regression to identify patients with MASH.
- More studies are needed to refine ML models for further evaluation including external validation.

LIMITATIONS

- A cross-sectional, observational study design has inherent internal validity limitations.
- Only non-institutionalized patients were included.

REFERENCES

1. Edlin, B. R., Eckhardt, B. J., Shu, M. A., Holmberg, S. D., & Swan, T. (2015). Toward a more accurate estimate of the prevalence of hepatitis C in the United States. *Hepatology*, *62*(5), 1353-1363.
2. Kowdley KV, Wang CC, Welch S, Roberts H, Brosgart CL. Prevalence of chronic hepatitis among foreign-born persons living in the United States by country of origin. Hepatology 2012;56:422-433.
3. Roberts, H., Kruszon-Moran, D., Ly, K. N., Hughes, E., Iqbal, K., Jiles, R. B., & Holmberg, S. D. (2016). Prevalence of chronic hepatitis in US households: National Health and Nutrition Examination Survey (NHANES), 1988-2012. *Hepatology*, *63*(2), 388-397.
4. WHO. Global health sector strategy on viral hepatitis 2016-2021; 2016. The Centers for Disease Control and Prevention's (CDC). Viral Hepatitis Surveillance United States, 2015. In; 2017.
5. Dinani AM, Kowdley KV, Noureddin M. Application of artificial intelligence for diagnosis and risk stratification in NAFLD and NASH: the state of the art. Hepatology. 2021;74:2233–40. 23.
6. National Health and Nutrition Examination Survey. 2017-2018 Data documentation, codebook, and frequencies: liver ultrasound transient elastography (LUX_J). 2020. [cited 2021 June]. Available from: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/ LUX_J.htm