RESEARCH PROPOSAL
Gastroenterology & Hepatology Research 2023;5(1):2. https://doi.org/10.53388/ghr2023-03-066

Classification of metabolic-associated fatty liver disease subtypes based on TCM clinical phenotype

Chenxia Lu[1, 2, 3], Hui Zhu[4], Mingzhong Xiao[1, 2, 3], Xiaodong Li[1, 2, 3*]

[1]Department of Hepatology, Hubei Key Laboratory of the theory and application research of liver and kidney in traditional Chinese medicine, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan 430061, China. [2]Affiliated Hospital of Hubei University of Chinese Medicine, Wuhan 430061, China. [3]Hubei Province Academy of Traditional Chinese Medicine, Wuhan 430074, China. [4]The Clinical Medical College of Traditional Chinese Medicine, Hubei University of Chinese Medicine, Wuhan 430065, China.

[*]**Corresponding to:** Xiaodong Li, Department of Hepatology, Hubei Key Laboratory of the theory and application research of liver and kidney in traditional Chinese medicine, Hubei Provincial Hospital of Traditional Chinese Medicine, No.4 Garden Hill, Wuchang District, Wuhan 430061, China. E-mail: lixiaodong555@126.com.

Author contributions

XD Li and MZ Xiao designed the study and revised the manuscript. CX Lu and H Zhu contributed significantly to the EMR data collection and analyses and wrote the manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by grants from the Key project Natural Science Foundation of Hubei Province (No. 2020CFA023); Project of the State Administration of Traditional Chinese Medicine (No Z155080000004): Key Laboratory of Liver and Kidney Treatment of Chronic Liver Diseases.

Peer review information

Gastroenterology & Hepatology Research thanks all anonymous reviewers for their contribution to the peer review of this paper.

Abbreviations

MAFLD, metabolic-associated fatty liver disease; NAFLD, non-alcoholic fatty liver disease; HEMnet, heterogeneous medical record network; EMR, electronic medical records; HCPSAS, Human-machine cooperative phenotypic spectrum annotation system; ICD-10, international classification of diseases; TCM, traditional Chinese medicine; t-SNE, t-distribution random neighbor embedding probability; WBC, white blood cell; RBC, red blood cell; PLT, platelet; HGB, hemoglobin; TP, total protein; ALB, albumin; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, R-glutamyl transferase; ALP, alkaline phosphatase; CHOL, serum cholesterol; HDL, high density cholesterol; LDL, low density cholesterol; sdLDL, small and dense low density lipoprotein; TG, triglyceride; FBG, fasting blood glucose; PLT serum platelet count; CRP, C-reactive protein; HbAlc, hemoglobin A1c; HOMA-IR, and insulin resistance index.

Citation

CX Lu, H Zhu, MZ Xiao, XD Li. Classification of metabolic-associated fatty liver disease subtypes based on TCM clinical phenotype. Gastroenterol Hepatol Res. 2023;5(1):2. doi: 10.53388/ghr2023-03-066.

Executive editor: Miao Peng.

Received: 09 March 2023; **Accepted:** 24 March 2023; **Available online:** 29 March 2023.

Abstract

Objective: To classify the subtypes of metabolic-associated fatty liver disease (MAFLD) and provide new insights into the heterogeneity of MAFLD.

Methods: Electronic medical records (EMR) of MAFLD diagnosed in accordance with the diagnostic criteria of Hubei Provincial Hospital of Traditional Chinese Medicine from 2016–2020 were included in the study. for physical annotation, and the data on each clinical phenotype was normalized according to corresponding aspirational standards. The MAFLD heterogeneous medical record network (HEMnet) was constructed using sex, age, disease diagnosis, symptoms, and Western medicine prescriptions as nodes and the co-occurrence times between phenotypes as edges. K-means clustering was used for disease classification. Relative risk (RR) was used to assess the specificity of each phenotype. Statistical methods were used to compare differences in laboratory indicators among subtypes.

Results: A total of patients (12,626) with a mean age of 55.02 ($\pm$14.21) years were included in the study. MAFLD can be divided into five subtypes: digestive diseases (C0), mental disorders and gynecological diseases (C1), chronic liver diseases and decompensated complications (C2), diabetes mellitus and its complications (C3), and immune joint system diseases (C4).

Conclusions: Patients with MAFLD experience various symptoms and complications. The classification of MAFLD based on the HEMnet method is highly reliable.

Keywords: metabolic-associated fatty liver disease; electronic medical records; disease classification; heterogeneous medical record network; disease heterogeneity

Gastroenterology & Hepatology Research 2023;5(1):2. https://doi.org/10.53388/ghr2023-03-066

Background

In 2020, metabolic-associated fatty liver disease (MAFLD) was adopted as a new term to better describe the pathogenesis of non-alcoholic fatty liver disease (NAFLD) [1]. Subsequently, the expert consensus on the name change to MAFLD was published in the Asia-Pacific region, the Middle East, and North Africa, and new criteria for the diagnosis of MAFLD were proposed [2, 3]. This nominal change emphasizes that multisystem metabolic dysfunction may be an important driver of the occurrence and adverse outcomes of MAFLD in the liver and other organs.

Although "MAFLD" is the only umbrella term for this disease, further research on the classification of MAFLD is needed in order to identify pathophysiological pathways that have similar clinical presentations but different clinical outcomes. Understanding the classification of MAFLD would be of great value for the diagnosis and treatment of this disease [4].

In this study, real-world clinical electronic medical records (EMRs) of MAFLD diagnosed using the diagnostic criteria of traditional Chinese medicine were used as the data source, and a Human-machine Cooperative Phenotypic Spectrum Annotation System [5] was used to obtain high-quality MAFLD clinical characterization datasets. The heterogeneous medical record network (HEMnet) was used to conduct an in-depth analysis of the dataset, extract the optimal combination module corresponding to each clinical element of MAFLD, and establish the disease classification subtype and phenotypic-molecular network association of MAFLD.

Methods

Screening and exclusion criteria for EMRs

EMRs of discharge diagnoses that met the MAFLD diagnostic criteria of Hubei Provincial Hospital of Traditional Chinese Medicine (TCM) from 2016–2020 were included. The diagnostic criteria for MAFLD were as follows: presence of fatty liver indicated by liver biopsy histology, or imaging; or blood biomarker examination indicating the presence of fatty liver combined with a diagnosis of obesity, type 2 diabetes, or metabolic dysfunction [1].

Exclusion criteria for EMRs were as follows: records that did not include a diagnosis of MAFLD, incomplete records, and MAFLD medical records in paper form.

TCM clinical characterization labeling

EMR data were imported into the Human-machine Cooperative Phenotypic Spectrum Annotation System (http://www.tcmai.org/login, HCPSAS) for physical annotation, including Western medicine disease diagnosis, TCM disease diagnosis, and TCM clinical symptoms. The basic process of annotation is "batch intelligent iteration - a small number of manual annotations-manual re-audits."

Regularization of clinical representations

Various clinical phenotypes, including diagnosis of diseases using TCM and Western medicine, symptoms of TCM, drug prescription, and physical and chemical indexes were standardized. Processing included original data extraction, data cleaning, synonym replacement, and feature dimension reduction processing.

The principles of dimensionality reduction for Western medicine diagnosis features are based on the International Classification of Diseases (ICD-10, Domestic Clinical Edition), Internal Medicine (9th edition), and Practical Internal Medicine (15th edition). The principles of standardized treatment of TCM symptoms refer to the diagnosis of TCM, differential diagnosis of TCM symptoms, terminology specification of common clinical symptoms of TCM, and differential diagnosis of symptoms. The principles of standardized treatment with Western medicine refer to the Pharmacopoeia of the People's Republic of China (2015), the National Catalogue of Essential Medicines (2018 Edition), and the National Catalogue of Pharmaceutical Insurance (2020).

MAFLD heterogeneous medical record network (HEMnet) and network embedding

A HEMnet based on MAFLD was constructed with phenotypes such as gender, age, disease diagnosis, symptoms, western prescription medication (skin, eye, ear, nose, and throat diagnosis, biologics medication (saline and glucose solution excluded) as nodes; and the co-occurrence times between phenotypes as edges [6]. In this study, four different side categories were used to create HEMnet: namely protein-protein interactions, herb targets, herb–symptom dictionary, and MAFLD EMRs. The first three categories were obtained from an external biomedical database, whereas the last category was extracted directly from MAFLD clinical electronic medical records.

To further examine the relationship between nodes in the network, all nodes in the HEMnet were first embedded into the 500-dimensional vector space using the ProSNet algorithm. Using a heterogeneous network as the input, the low-dimensional vector representation of each node was optimized.

Feature matrix clustering

In this study, we used the k-means clustering algorithm to cluster the MAFLD electronic medical record feature matrix. Because the individuals in each cluster were highly homogenous, the individuals in different cluster sets were highly heterogeneous. Therefore, we can consider each cluster module as a subset of the HEMnet.

Results

Demographic information

A total of 12,626 patients with MAFLD were included in the study, with an average age of 55.02 (±14.21) years, including 7,067 men with an average age of 51.9 (±14.49) years. Furthermore, there were 5,595 women with a mean age of 58.99 (±12.8) years (Figure 1).

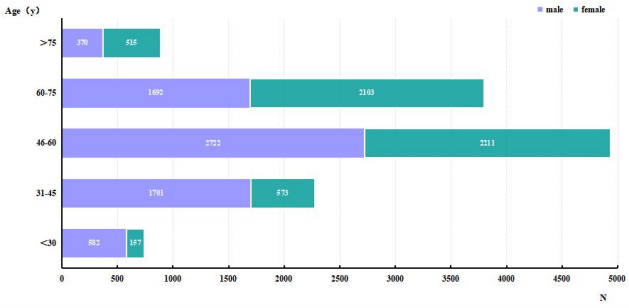

Figure 1 Characteristics of 12,626 MAFLD patients

Assessment of subtype effect and overall distribution characteristics

High-quality and reliable cluster analysis is essential to identifying disease subtypes. In this study, Dunn's index, Kullback-Leibler (K-L) divergence, and t-distribution stochastic neighbor embedding (t-SNE) were introduced to evaluate the clustering effect of disease classification.

Dunn's index considers not only the distance between the centroid of a module subtype and nodes within the subtype (intraclass distance), but also the distance between the centroid of two clustering modules (interclass distance). Its goal is to minimize the intraclass distance while satisfying the maximum interclass distance, and the index score is the ratio of the minimum interclass distance to the maximum intraclass distance. For the same number of classes, the larger the Dunn's index, the closer the class structure and the greater the separation between classes, which should correspond to better-quality clustering results. However, because Dunn's index is very sensitive to the classification of high-dimensional data, even under the standard class number, it cannot fully reflect the quality difference and class number estimation of the clustering results.

To improve the classification accuracy of MAFLD disease subtypes, the similarity of classification variables (TCM symptoms,

Gastroenterology & Hepatology Research 2023;5(1):2. https://doi.org/10.53388/ghr2023-03-066

comorbidities, age stratification, and Western medicine prescriptions) in the HEMnet model should be calculated during classification and used as the classification basis. If the two models are similar, their similarity to the same observed quantity will also be similar. Therefore, this study used the K-L divergence to measure the differentiation ability of the K-means clustering model to divide disease subtypes and evaluate the degree of difference between different subtypes.

The clustering analysis of this study shows that when the cluster number was 5, the Davies-Bouldin index had the lowest value (1.6245), the K-L divergence had the highest value (2.1403), and the Dunn index ranked third globally (0.0643). At this point, the clustering effect was optimal (Table 1).

Table 1 Cluster quality assessment index of each subtype

k	dunn_index	Silhouette score	davies_bouldin score	calinski_harabasz score	Kullback Leibler divergence
2	0.0769	0.2610	2.2342	1490.0455	2.1339
3	0.0744	0.2764	1.8164	1504.0192	2.1259
4	0.0571	0.1565	1.7979	1508.1985	2.1221
5	0.0643	0.1690	1.6245	1447.1465	2.1403
6	0.0467	0.1060	1.9717	1287.1809	2.1347
7	0.0485	0.1065	2.1260	1180.8636	2.1219
8	0.0381	0.0716	2.1691	1078.8420	2.1347
9	0.0381	0.0765	2.1058	991.5843	2.1333
10	0.0393	0.0788	2.0416	917.4416	2.1363

t-SNE dimension reduction visualization

In this study, the dimensionality reduction method based on K-L divergence and the t-distribution random neighbor embedding probability (t-SNE) dimensionality reduction analysis was used to process the dimensionality reduction visualization model of high-dimensional clinical phenotype data. The relative positions of the MAFLD patients included in the analysis were mapped to the low-dimensional space, and the internal structure of each subtype was presented. The basic principle is to convert the distance into the probability distribution between people using the radial basis function; that is, people with similar characteristics are more likely to become adjacent nodes. During the dimensionality reduction process, the objective function with K-L divergence as the core can make the clinical phenotype data after dimensionality reduction maintain the relative position of the population in the high-dimensional space. The visualization data after dimensionality reduction are shown in Figure 2. Differentiation among the five subtypes was apparent. There were 8,351 cases of the C0 subtype. The C0 subtype accounted for 66.14% of MAFLD cases which made C0 the most common subtype. There were 1,694 cases of the C1 subtype. There were 384 cases of the C2 subtype, and patients with the C2 subtype were typically older and had longer hospital stays. The C3 subtype was the second most common subtype. There were 1,862 cases of the C3 subtype accounting for 14.75% of MAFLD cases. Last, 335 patients had the C4 subtype making this the least common subtype.

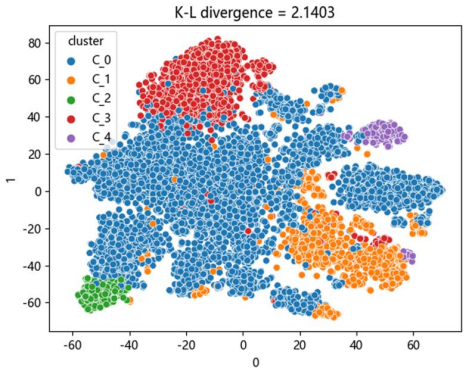

Figure 2 Visualization of t-SNE dimension reduction in MAFLD subtypes

Distribution of co-morbidity characteristics of MAFLD subtypes

Cij > 20, RR > 1.5, *P* < 0.05, and the proportion of the disease in the subtype accounting for more than 50% of the disease frequency was set as the screening condition for the characteristics of the subtype. The results showed that 33 MAFLD co-morbidities were associated with the C0 subtype, mainly digestive system diseases, including gastritis, intestinal polyps, esophagitis, cholecystitis, peptic ulcer, enteritis, helicobacter pylori infection, gastric polyps, diarrhea, pancreatitis, and gastroesophageal reflux disease (Table 2).

The C1 subtype was mainly associated with mental disorders and gynecological diseases, including breast hyperplasia, facial paralysis, uterine fibroids, sleep disorders, breast cancer, anxiety, synovitis, endometrial thickening, anxiety, and depression. Moreover, it was associated with limb joint diseases, such as lumbar spine disease and knee degenerative changes (Table 3).

The C2 subtype was mainly associated with chronic liver disease and decompensated complications, including cirrhosis, viral hepatitis, cholestasis, hypoproteinemia, peritonitis, ascites, liver failure, portal hypertension, liver cancer, esophageal and gastric varices, hypersplenism, hepatic encephalopathy, hepatitis C, gastrointestinal bleeding, hepatic occupying lesions, and hepatitis (Table 4).

Type 2 diabetes mellitus and its complications were the main comorbidities in patients with the C3 subtype, including type 2 diabetes, diabetic complications, hyperlipidemia, atherosclerosis, type 2 diabetic peripheral neuropathy, type 2 diabetic vascular disease, type 2 diabetic retinopathy, gout, diabetic nephropathy, and ketosis (Table 5).

The C4 subtype was associated with rheumatoid immune system disorders, including arthritis, osteoporosis, lumbar spine disease, joint effusion, gout, connective tissue disease, Sjogren's syndrome, and systemic lupus erythematosus (Table 6).

Table 2 Features of the C0 subtype co-morbidity

feature	cij	rate (%)	RR	chiq_p
Irritable bowel syndrome	66	95.65	11.34387447	2.06E-07
Intestinal diverticula	40	95.24	10.28275779	6.5693E-05
diarrhea	96	94.12	8.274258025	2.03627E-09
Myelodysplastic syndrome	26	92.86	6.672552553	0.002784286
purpura	34	91.89	5.821329806	0.000917173
Adrenal hyperplasia	47	90.38	4.833574181	0.000214036
Intestinal polyp	362	87.65	3.752924978	5.87716E-21
pancreatitis	91	87.50	3.611864407	3.80328E-06
IgA nephropathy	42	87.50	3.596461668	0.00173069
vertigo	83	87.37	3.566249395	1.14107E-05

Gastroenterology & Hepatology Research 2023;5(1):2. https://doi.org/10.53388/ghr2023-03-066

feature	cij	rate (%)	RR	chiq_p
Barrett's esophagus	69	87.34	3.55330838	6.48607E-05
enteritis	153	86.93	3.450237067	4.37206E-09
Intestinal obstruction	25	86.21	3.206071343	0.022253411
Gastroesophageal reflux disease	91	85.85	3.128813559	1.66416E-05
Somatization disorder	91	85.85	3.128813559	1.66416E-05
Diffuse toxic goiter	35	85.37	2.99452862	0.009175316
Cerebral artery stenosis	133	85.26	2.991926524	3.84381E-07
Angina pectoris	289	84.75	2.911204988	1.81442E-13
Helicobacter pylori infection	150	84.75	2.877697842	1.38354E-07
cholecystitis	275	83.84	2.712556188	6.70984E-12
Cerebral ischemia	386	83.19	2.607629533	2.64101E-15
hemorrhoids	89	83.18	2.547627693	0.0001842
Gastrointestinal dysfunction	42	82.35	2.395956192	0.014228971
headache	83	81.37	2.248669569	0.001099228
Gastric polyp	125	81.17	2.224863134	7.34009E-05
nephritis	168	80.77	2.173652695	6.95005E-06
constipation	38	80.85	2.166726813	0.03276294
Space occupying lesion of lung	38	80.85	2.166726813	0.03276294
gastritis	1662	78.88	2.13849671	9.43064E-42
prediabetes	362	80.27	2.131205689	1.08213E-10
Sleep apnea hypopnea syndrome	49	80.33	2.096753794	0.018922659
pneumonia	218	80.15	2.09520882	8.03299E-07
pharyngitis	72	80.00	2.056770141	0.005299988

Table 3 Features of the C1 subtype co-morbidity

feature	cij	rate (%)	chiq_p	RR
Lumbar disease	296	21.64	3.47396E-21	1.781899241
Mammary hyperplasia	62	28.44	5.19852E-11	2.564798838
Facial paralysis	50	74.63	3.60808E-49	18.98048476
Uterine myoma	48	27.27	5.59571E-08	2.420011806
Sleep disorder	47	20.61	0.001292016	1.675735615
Nephrotic syndrome	46	21.60	0.00041181	1.777573542
Breast cancer	38	40.43	1.24142E-14	4.379068983
Thyroid cyst	38	25.17	2.03135E-05	2.17015808
rhinitis	30	23.81	0.000581798	2.016676505
Varicose veins	29	20.28	0.015450018	1.641645436
Lumbar muscle strain	28	53.85	1.02189E-17	7.52892562
Anxious state	28	25.45	0.000198822	2.203587986
synovitis	27	26.73	8.07489E-05	2.354606082
Increased endometrium thickness	25	83.33	4.72117E-28	32.26682409
Soft tissue injury	25	65.79	2.38688E-21	12.41031696
neuritis	23	43.40	1.38952E-10	4.947579693
Anxiety-depressive state	23	25.27	0.000865674	2.182755747
Degenerative changes of knee joint	23	23.96	0.002350201	2.033251929
Pleural effusion	22	27.85	0.000159865	2.490772386
Cerebral hemorrhage	22	25.29	0.00111485	2.184215784

Table 4 Features of the C2 subtype co-morbidity

Gastroenterology & Hepatology Research 2023;5(1):2. https://doi.org/10.53388/ghr2023-03-066

feature	cij	rate (%)	chiq_p	RR
Cirrhosis of liver	344	52.60	2.2251E-308	35.37674731
cholestasis	118	62.43	2.2251E-308	52.98400822
peritonitis	107	84.25	2.2251E-308	170.5591146
ascites	101	67.33	2.2251E-308	65.71226616
Liver failure	90	90.00	2.2251E-308	286.921875
Portal hypertension	79	75.96	2.2251E-308	100.7414583
Liver cancer	78	58.65	2.5679E-304	45.21193182
Esophageal fundus varices	60	69.77	2.3777E-281	73.56971154
Hypersplenism	57	74.03	1.316E-284	90.85859375
Hepatic encephalopathy	51	64.56	2.57E-219	58.06752232
Hepatic space occupying lesion	23	57.50	8.92147E-86	43.13204657
hepatitis	21	72.41	4.18748E-100	83.68554688

Table 5 Features of the C3 subtype concomitant disease

feature	cij	rate (%)	chiq_p	RR
Complications of diabetes	1467	89.89	2.2251E-308	51.39728542
Type 2 diabetic peripheral neuropathy	815	93.46	2.2251E-308	82.65645316
Type 2 diabetic vascular disease	745	95.88	2.2251E-308	134.5861305
Type 2 diabetic retinopathy	540	91.68	2.2251E-308	63.70766567
Diabetic nephropathy	321	88.67	2.2251E-308	45.26006654
ketoacidosis	263	90.38	1.1596E-296	54.29898726
cataract	88	51.16	6.74583E-42	6.05616081

Table 6 Features of the C4 subtype complications

feature	cij	chiq_p	RR
arthritis	283	2.2251E-308	42.55386592
osteoporosis	208	2.2251E-308	25.35357763
gallstone	83	3.09452E-06	1.609531097
Lumbar disease	61	1.07675E-05	1.712366248
Effusion of knee joint	60	7.3094E-215	33.3541384
Gout disease	56	5.60579E-07	1.902421227
Connective tissue disease	53	9.8792E-196	35.3553867
Kidney stone	49	0.001545148	1.545819484
Electrolyte metabolism disorder	45	7.99844E-21	3.804216246
Renal cyst	42	0.000778349	1.65516777
Sjogren's syndrome	34	4.0065E-123	34.65124378
Pulmonary infection	32	1.45268E-09	2.835907419
anemia	30	2.38201E-07	2.518733563
Urinary tract infection	30	0.002005475	1.74159267
Systemic lupus erythematosus	28	2.32088E-74	21.40223881
hypothyroidism	26	4.5599E-05	2.197991609

Comparison of clinical adverse events for each subtype

Cardiovascular events, malignancies, and organ failure were defined as clinical adverse events in MAFLD cases. Cardiovascular events include coronary heart disease and stroke events involved in the diagnosis of patients using Western medicine. Coronary events included acute myocardial infarction, ischemic cardiac arrest, death from chronic coronary heart disease, and coronary stent placement or coronary artery bypass grafting. Stroke events included subarachnoid hemorrhage, intracerebral hemorrhage, cerebral infarction, cerebral embolism, and unclassified stroke. Organ failure included failure of all organs or systems involved in the diagnosis, namely respiratory failure, heart failure, kidney failure, and liver failure.

The results showed that the C2 subtype exhibited the highest rate of malignancy and organ failure, and the C3 subtype exhibited the highest rate of cardiovascular events (Table 7).

Table 7 Comparison of clinical adverse events and hospitalized day of each subtype

	C0	C1	C2	C3	C4
	N = 8351	N = 1694	N = 384	N = 1862	N = 335

Gastroenterology & Hepatology Research 2023;5(1):2. https://doi.org/10.53388/ghr2023-03-066

	C0 N = 8351	C1 N = 1694	C2 N = 384	C3 N = 1862	C4 N = 335
Cardiovascular event	2476 (29.65)	413 (24.38)	19(4.95)	580 (31.15)	65 (19.40)
Malignant tumor	577 (6.91)	292 (17.24)	103 (26.82)	23(1.24)	7 (2.09)
Organ failure	547 (6.55)	116 (6.85)	248 (64.58)	58(3.11)	19(5.67)
Average length of stay	9.36 ± 5.58	10.34 ± 6.08	19 ± 13.61	9.44 ± 4.12	8.79 ± 3.16

Comparison of laboratory test indices of each subtype

The laboratory examination results of MAFLD patients at first admission were compared among the subtypes. As the C0 subtype accounted for the largest proportion of patients, the test results of each subtype were statistically compared with those of the C0 subtype, as shown in Table 8.

The results showed that the C2 subtype showed clear characteristics of reduced white blood cell (WBC), red blood cell (RBC), platelet (PLT), hemoglobin (HGB) counts, abnormal liver function, and decreased protein level. Namely, serum levels of WBC, RBC, PLT, hemoglobin (HGB), total protein (TP), and albumin (ALB) were significantly lower than those of other subtypes ($P < 0.001$, except WBC). Serum levels of aspartate aminotransferase (AST), alanine aminotransferase (ALT), R-glutamyl transferase (GGT), and alkaline phosphatase (ALP) were higher than those of other subtypes ($P < 0.001$, except ALT). Moreover, the C2 subtype also had clear hypolipidemia, that is, serum cholesterol (CHOL), high density cholesterol (HDL), low density cholesterol (LDL), small and dense low density lipoprotein (sdLDL), and triglyceride (TG) levels that were significantly lower than those of other subtypes ($P < 0.001$).

The C3 subtype showed clear abnormal blood glucose and severe insulin resistance; that is, the fasting blood glucose (FBG), hemoglobin A1c (HbAlc), and insulin resistance index (HOMA-IR) were significantly higher than those of other subtypes ($P < 0.001$, except HbAlc). The C3 subtype also exhibited hyperlipidemia. Compared with other subtypes, high cholesterol (CHOL), triglyceride (TG), and sdLDL levels were consistent with the features of this subtype.

The C4 subtype showed significantly higher levels of inflammatory characteristics, including PLT, C-reactive protein (CRP), and WBC ($P < 0.001$, excluding WBC).

Table 8 Comparison of biochemical laboratory indicators among each subtype

laboratory indicators	C0	C1	C2	C3	C4
WBC (*10^9/L)	6.62 ± 2.41	6.66 ± 2.72	5.68 ± 3.68	6.37 ± 1.81***	6.73 ± 2.69
RBC (*10^12/L)	4.57 ± 0.61	4.5 ± 0.62**	3.41 ± 0.9***	4.65 ± 0.57***	4.19 ± 0.52***
PLT (*10^9/L)	213.46 ± 77.42	217.29 ± 66.64	109.6 ± 75.62***	204.81 ± 57.35***	231.25 ± 78.8**
HGB (g/L)	138.84 ± 18.35	136.76 ± 18.87**	110.59 ± 28.05***	141.36 ± 16.67***	124.81 ± 16.49***
CRP (mg/L)	10.12 ± 27.01	8.08 ± 20.97	17.58 ± 34.87	5.26 ± 15.22***	23.88 ± 39.33***
AST (U/L)	28.04 ± 37.94	24.77 ± 18.11***	81.46 ± 108.55***	23.58 ± 20.95***	22.98 ± 16.45***
ALT (U/L)	38.14 ± 64.52	32.28 ± 30.38***	63.47 ± 115.25	31.1 ± 31.88***	26.29 ± 23.27***
GGT (U/L)	48.97 ± 84.68	44.12 ± 83.75	130.36 ± 201.51***	41.11 ± 63.45***	39.08 ± 57.99
ALP (U/L)	79.67 ± 39.92	81.22 ± 42.47	162.04 ± 106.23***	81.3 ± 30.88	87.55 ± 40.48*
UA (μmol/L)	394.51 ± 112.63	384.8 ± 109.73*	426.6 ± 191	366.63 ± 106.12***	334.05 ± 130.4***
ALB (g/L)	41.51 ± 4.71	41.19 ± 4.6	31.55 ± 7.07***	41.27 ± 3.81	39.16 ± 4.34***
TP (g/L)	69.89 ± 6.83	69.55 ± 6.75	64.07 ± 11.89***	69.2 ± 6.22**	69.19 ± 6.45
CHOL (mmol/L)	4.95 ± 1.22	5.05 ± 1.33	3.31 ± 1.58***	5.03 ± 1.32	4.72 ± 1.06**
HDL (mmol/L)	1.11 ± 0.29	1.11 ± 0.26	0.92 ± 0.41**	1.04 ± 0.27***	1.1 ± 0.3
LDL (mmol/L)	2.84 ± 0.84	2.92 ± 0.87**	2.23 ± 1.01***	2.87 ± 0.86	2.76 ± 0.75
sdLDL (mmol/L)	1.03 ± 0.4	1.05 ± 0.5	0.59 ± 0.49***	1.08 ± 0.38***	0.91 ± 0.34**
TG (mmol/L)	2.36 ± 2.16	2.31 ± 1.86	1.06 ± 0.67***	2.73 ± 2.85***	1.89 ± 1.29***
FBG (mmol/L)	6.31 ± 2.49	6.05 ± 2.44*	7.53 ± 8.82	11.11 ± 5.59***	5.65 ± 1.44***
HbAlc (%)	6.55 ± 1.58	6.38 ± 1.55	7.19 ± 2.77***	9.05 ± 2.28	6.56 ± 1.4
IRI (uIU/mL)	14.43 ± 10.29	15.96 ± 11.82	14.84 ± 8.33	16.46 ± 18.72	23.95 ± 17.43
HOMAIR	4.82 ± 4.62	5.18 ± 4.43	4.81 ± 2.84	7.69 ± 8.95***	7.63 ± 7.13

Discussion

The development of new data-driven disease classifications that capture heterogeneous clinical manifestations has the potential to improve the understanding and clinical care of MAFLD and advance precision diagnosis and treatment. Our study used a data-driven, unsupervised clustering approach to find MAFLD subtypes and group MAFLD patients based on similarities between TCM clinical phenotypes derived from real-world unstructured clinical electronic medical records. We hypothesized that this approach may provide new insights into the etiology and defining characteristics of the disease.

A Human-machine Cooperative Phenotypic Spectrum Annotation System (http://www.tcmai.org/login, HCPSAS) can extract phenotypes from MAFLD clinical electronic medical records by combining a small number of manual annotations with large-scale active learning. It forms high-quality structured data quickly and iteratively to solve the problems of large amounts of clinical text, medical record processing, rich information granularity, and complex structure. Furthermore, the probability graph model and the network medical analysis methods were introduced to integrate the multidimensional and heterogeneous clinical characteristics of TCM symptoms, physical and chemical indices, prescription drugs, and disease genes. The HEMnet method was applied to conduct the disease classification process. The precise classification of disease syndromes from single TCM symptoms to integrated multidimensional clinical representations (symptoms, physical and chemical indicators, baseline

data) was performed. This enabled verification of the scientific nature and reliability of the analogical similarity between "disease groups" as the components of the disease syndrome classification model in terms of mathematical algorithms.

In this study, we showed that MAFLD symptoms are diverse, complications are complex, clinical characteristics of sex and age are variable, and the symptom-symptom, symptom-syndrome, and symptom-disease relationships are unclear and non-linear. Adverse clinical events were observed in the entire study population. Cardiovascular events accounted for 28.14%, malignant tumors for 7.94%, and organ failure for 7.83% of events. The causal relationship between MAFLD and cardiovascular events and malignancies was not analyzed in this study. Nonetheless, clinicians managing patients with MAFLD should not only focus on liver disease but also consider the increased risk of cardiovascular disease and tumor, and conduct early, aggressive risk factor modification.

In this study, 12,626 patients with MAFLD were divided into five subtypes with typical characteristics, including patient age, co-morbidities, TCM symptoms, prescriptions of Western medicine, and mapping gene targets in the medical structure network. Our data suggests that attention should be paid to MAFLD-related digestive diseases, including gastritis and enteritis.

In addition to the adverse outcomes of chronic liver disease and complications associated with diabetes, it is worth noting that menstruation, reproductive status, and sex hormone factors are closely related to the occurrence of MAFLD [7]. In this study, women with the C1 subtype showed prominent manifestations of abnormal menstruation and mental disorders. Irregular menstrual cycles could increase the risk of MAFLD, regardless of reproductive age or premenopausal status [8, 9]. Recently, neurological disorders and behavioral changes such as depression and anxiety have been reported as common manifestations of MAFLD [10, 11]. These new clinical representations also highlight the heterogeneity and pathogenesis of MAFLD.

In this study, a small number of patients also had knee osteoarthritis, which is related to obesity and metabolic disorders. In a health survey of 17,476 patients with MAFLD in South Korea, MAFLD was significantly associated with knee osteoarthritis [12]. Therefore, we should simultaneously manage and treat C4 subtypes with such clinical characteristics.

Conclusions

The heterogeneity among patients with MAFLD is particularly significant due to genetic, environmental, lifestyle and other factors. These patients have great differences in laboratory indicators, symptoms, comorbidities and post-treatment responses, which is an important reason for the obstruction of clinical diagnosis and treatment of the disease and the study of data models. In this study, probability graph model and network medical analysis method were introduced to integrate TCM symptoms, physical and chemical indicators, disease genes and other multi-dimensional clinical characteristics, and HEMnet method was applied to study the disease classification of MAFLD from retrospective electronic medical record data, providing methods for exploring the complexity of diseases, disease classification and individualized diagnosis and treatment.

References

1. Eslam M, Newsome PN, Sarin SK, et al. A new definition for metabolic dysfunction-associated fatty liver disease: An international expert consensus statement. *J Hepatol* 2020;73(1):202–209. Available at: http://doi.org/10.1016/j.jhep.2020.03.039
2. Eslam M, Sarin SK, Wong VW-S, et al. The Asian Pacific Association for the Study of the Liver clinical practice guidelines for the diagnosis and management of metabolic associated fatty liver disease. *Hepatol Int* 2020;14(6):889–919. Available at: http://doi.org/10.1007/s12072-020-10094-2
3. Shiha G, Alswat K, Al Khatry M, et al. Nomenclature and definition of metabolic-associated fatty liver disease: a consensus from the Middle East and north Africa. *T Lancet Gastroenterol Hepatol* 2021;6(1):57–64. Available at: http://doi.org/10.1016/S2468-1253(20)30213-2
4. Xue R, Fan JG. Brief introduction of an international expert consensus statement: A new definition of metabolic associated fatty liver disease. *J Clinical Hepatol* 2020,36(6):1224–1227. Available at: https://doi.org/10.3969/j.issn.1672-5069.2020.03.039
5. Zou Q, Yang K, Shu Z, et al. Phenonizer: A Fine-Grained Phenotypic Named Entity Recognizer for Chinese Clinical Texts. *Biomed Res Int* 2022;2022:3524090. Available at: http://doi.org/10.1155/2022/3524090
6. Huang EW, Wang S, Li B, et al. HEMnet: Integration of Electronic Medical Records with Molecular Interaction Networks and Domain Knowledge for Survival Analysis 2017. Available at: http://doi.org/10.1145/3107411.3107422
7. Stefan N, Cusi K. A global view of the interplay between non-alcoholic fatty liver disease and diabetes. *Lancet Diabetes Endocrinol* 2022;10(4):284–296. Available at: http://doi.org/10.1016/S2213-8587(22)00003-1
8. Kim W. Epidemiologic Landscape of Nonalcoholic Fatty Liver Disease Is Changed During Lifetime by Menstrual and Reproductive Status and Sex Hormonal Factors. *Clin Gastroenterol Hepatol* 2021;19(6):1114–1116. Available at: http://doi.org/10.1016/j.cgh.2020.10.054
9. Cho IY, Chang Y, Kang J-H, et al. Long or Irregular Menstrual Cycles and Risk of Prevalent and Incident Nonalcoholic Fatty Liver Disease. *J Clin Endocrinol Metab* 2022;107(6):e2309–2317. Available at: http://doi.org/10.1210/clinem/dgac068
10. Rafiei R, Bemanian M, Rafiei F, et al. Liver disease symptoms in non-alcoholic fatty liver disease and small intestinal bacterial overgrowth. *Rom J Intern Med* 2018;56(2):85–89. Available at: http://doi.org/10.1515/rjim-2017-0042
11. Moretti R, Caruso P, Gazzin S. Non-alcoholic fatty liver disease and neurological defects. *Ann Hepatol* 2019;18(4):563–570. Available at: http://doi.org/10.1016/j.aohep.2019.04.007
12. Han AL. Association between metabolic associated fatty liver disease and osteoarthritis using data from the Korean national health and nutrition examination survey (KNHANES). *Inflammopharmacol* 2021;29(4):1111–1118. Available at: http://doi.org/10.1007/s10787-021-00842-7