Table S1. Description of the final feature subset using IGSFS-CD (Subset evaluation with Decision Tree) for two datasets.

LC_dataset-1		**LC_dataset II**	
Selected features	**Descriptions**	**Selected features**	**Descriptions**
1. AGE	Age range according to hospital code by classified into 16 age groups.	1. AGE	Age range according to hospital code by classified into 16 age groups.
2. EXT	Extent of metastasis 1=In situ 2=Localized 3=Direct extension 4=Regional lymph nodes 5=Distant metastases 8=Not applicable 9= Unknown	2. EXT	Extent of metastasis 1=In situ 2=Localized 3=Direct extension 4=Regional lymph nodes 5=Distant metastases 8=Not applicable 9= Unknown
3. BASIC	Diagnostic methods 0=Death Certificate Only 1=History & Physical exam 2= Endoscopy & Radiology 3= Surgery & Autopsy 4= Specific Biochem/Immuno tests 5= Cytology or Hematology 6= Histology of Metastasis 7= Histology of Primary	3. SURG	Surgical treatment 1=yes 2=No 3= Unknown
4. BC006	Cholesterol is a compound of the sterol type found in most body tissues. A blood test for cholesterol measures total cholesterol that is carried in the blood by lipoproteins. 1= Normal (0-200 mg/dl) 3= (>200 mg/dl) 9= Unknown	4. IM011	Carbohydrate antigen (CA 19-9) is a protein commonly found on the surface of cancer cells. It is useful as a tumor marker to follow the response to cancer treatment, particularly in pancreatic cancer. 1= Normal (0-39 U/ml) 3= (>39 U/ml) 9= Unknown
5. BC009	Triglycerides are chemical compounds and the most common form of fat in the body. Each triglyceride molecule contains a glyceride that is bound to three molecules of free fatty acids, called glycerol. The triglyceride test measures the level of triglycerides in the blood. 1= Normal (0-150 mg/dl) 3= (>150 mg/dl) 9= Unknown	5. 60_list	Cognition and Perception of pain by Gordon's functional health patterns 1=yes 2=No 9= Unknown
6. T	Primary tumor (T)	6. T	Primary tumor (T)
7. M	Distant metastasis (M)		

Table S2. Description of the final feature subset using IGSFS-CD (Subset evaluation with Naïve Bayes) for two datasets.

LC_dataset-1		**LC_dataset II**	
Selected features	**Descriptions**	**Selected features**	**Descriptions**
1. MOR	The morphology of the biopsy 8000 = Neoplasm, malignant 8010 = Carcinoma, NOS 8140 = Adenocarcinoma, NOS 8170 = Hepatocellular Carcinoma, NOS 8500 = Infiltrating duct carcinoma, NOS	1. BASIC	Diagnostic methods 0 = Death Certificate Only, 1 = History & Physical exam, 2 = Endoscopy & adiology, 3 = Surgery & Autopsy, 4 = Specific Biochem/Immuno tests, 5 = Cytology or Hematology, 6 = Histology of Metastasis, 7 = Histology of Primary
2. LATER	Laterality of tumor site such as right side or left side 0=Not a paired site 1 = Right: primary, 2 = Left: primary 3= One side, right or left unspecified, 4 = Bilateral, 9 = Pair site, laterality unknown	2. LATER	Laterality of tumor site such as right side or left side 0 = Not a paired site 1 = Right: primary, 2 = Left: primary, 3= One side, right or left unspecified, 4 = Bilateral, 9 = Pair site, laterality unknown
3. GRAD	Tumor Grade of the biopsy 1 =Well differentiated, 2 = Moderately differ'd, 3 = Poorly differ'd, 4 = Undifferentiated, 6 = Positive B-cell, 9 = Unknown	3. MET	Place of metastasis 0 = None, 1 = Lymph node, 2 = Bone, 3 = Liver, 4 = Lung / Pleura, 5 = Brain, 6 = Peritoneum, 7 = Multiple sites, 8 = Other, 9 = Unknown
4. EXT	Extent of metastasis 1 = In situ, 2=Localized, 3=Direct extension, 4=Regional lymph nodes, 5=Distant metastases, 8=Not applicable, 9= Unknown	4. EXT	Extent of metastasis 1 = In situ, 2 = Localized, 3=Direct extension, 4 = Regional lymph nodes, 5 = Distant metastases, 8 = Not applicable, 9 = Unknown
5. T	Primary tumor (T)	5. T	Primary tumor (T)
6. PDX2	ICD-10-CM Code of Principal Diagnosis such as C22.0 = Liver cell carcinoma, C22.2 = Hepatoblastoma, K76.9= Liver disease	6. PDX2	ICD-10-CM Code of Principal Diagnosis such as C22.0 = Liver cell carcinoma, C22.2 = Hepatoblastoma, K76.9= Liver disease
7. SDX6	ICD-10-CM Code of Secondary Diagnosis such as Z51.1=Chemotherapy session for neoplasm, K76.8= Other specified diseases of liver	7. SDX3	ICD-10-CM Code of Secondary Diagnosis such as Z51.1= Chemotherapy session for neoplasm

Table S2 Cont.

LC_dataset-1		LC_dataset II	
Selected features	**Descriptions**	**Selected features**	**Descriptions**
8. SDX9	ICD-10-CM Code of Secondary Diagnosis	8. CHEM	Chemical treatment 1 = yes, 2 = No, 3 = Unknown
9. PROC1	ICD-9-CM Code of Procedure such as 50.3 = Lobectomy of liver 99.25 = Injection or infusion of cancer chemotherapeutic substance	9. RADI	Radiation therapy 1 = yes, 2 = No, 3 = Unknown
10. PROC2	ICD-9-CM Code of Procedure	10. N	Regional lymph nodes (N)
11. BC005	Uric acid is produced by the breakdown of purines in cells of the body. In term of cancer treatment, a blood test for uric acid is useful to monitor uric acid levels when undergoing chemotherapy or radiation treatment. 1 = Normal (Male 2.4-5.7 mg/dl, Female 3.4-7.0 mg/dl) 2 = (Male <2.4 mg/dl, Female <3.4 mg/dl), 3 = (Male >5.7 mg/dl, Female>7.0) 9 = Unknown	11. BC009	Triglycerides are chemical compounds and the most common form of fat in the body. Each triglyceride molecule contains a glyceride that is bound to three molecules of free fatty acids, called glycerol. The triglyceride test measures the level of triglycerides in the blood. 1= Normal (0-150 mg/dl) 3= (>150 mg/dl) 9= Unknown
12. BC006	Cholesterol is a compound of the sterol type found in most body tissues. A blood test for cholesterol measures total cholesterol that is carried in the blood by lipoproteins. 1 = Normal (0-200 mg/dl), 3 = (>200 mg/dl), 9 = Unknown	12. BC015	It is known as serum glutamic-oxaloacetic transaminase (SGOT). AST is an enzyme mostly found in the heart and liver. AST test is useful for detecting or monitoring liver damage. 1 = Normal (0-32 U/L Female, 0-40 U/L Male), 3 = (>32 U/L Female, >40 U/L Male), 9 = Unknown
13. IM001	AFP is a protein produced by the liver when its cells are regenerating. It is normally used as a tumor marker to detect and diagnose cancers of the liver. 1 = Normal (0.0-13.60 ng/ml), 3 = (>13.60 ng/ml), 9 = Unknown	13. IM008	It is an indicator of the amount of cancer, size of tumor, and cancer staging. CEA is primarily used to monitor cancer treatment, including response to therapy and recurrence. 1= Normal (0-5 ng/ml), 3 = (>5 ng/ml) 9 = Unknown

Table S2 Cont.

LC_dataset-1		LC_dataset II	
Selected features	**Descriptions**	**Selected features**	**Descriptions**
14. IM008	It is an indicator of the amount of cancer, size of tumor, and cancer staging. CEA is primarily used to monitor cancer treatment, including response to therapy and recurrence. 1 = Normal (0-5 ng/ml), 3 = (>5 ng/ml), 9 = Unknown	14. 57_list	Cognition and perception of visibility by Gordon's functional health patterns. 1 =Normal vision, 2 = Shortsighted, 3 = Longsighted, 4 = Wear glasses, 5 = Other, 9 = Unknown