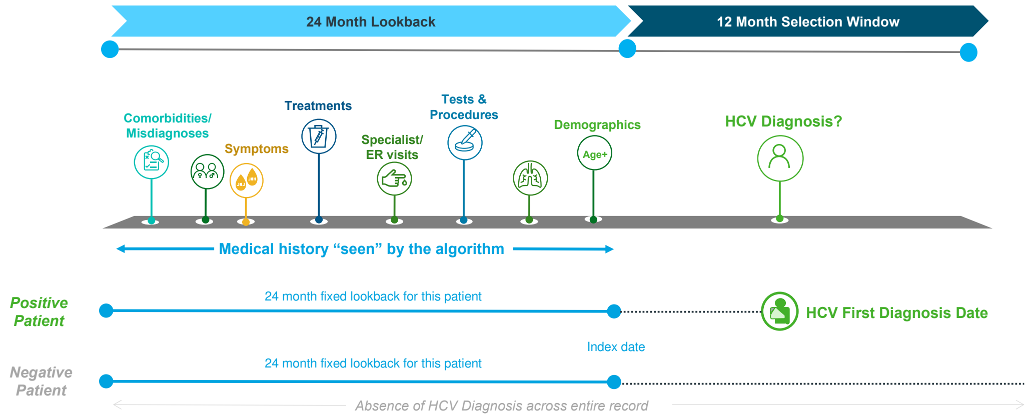
Supplementary Information for “Finding undiagnosed patients with Hepatitis C Virus: an application of artificial intelligence to US ambulatory electronic medical records”

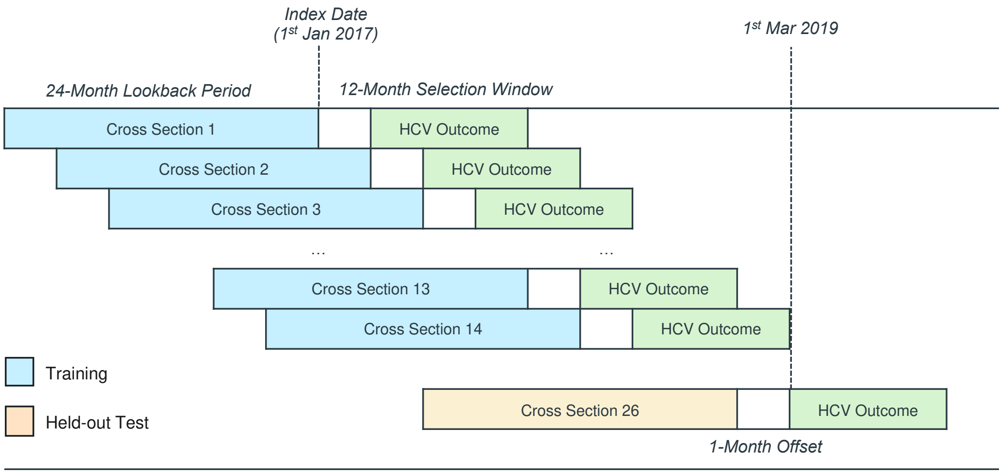
1 Supplementary Methods

Figure S 1 Study design: cross-sectional approach.


24 Month Lookback
12 Month Selection Window
Comorbidities/ Misdiagnoses
Symptoms
Treatments
Specialist/ ER visits
Tests & Procedures
Demographics
Age+
HCV Diagnosis?
Medical history "seen" by the algorithm
Positive Patient
24 month fixed lookback for this patient
HCV First Diagnosis Date
Index date
Negative Patient
24 month fixed lookback for this patient
Absence of HCV Diagnosis across entire record

The study was designed as a retrospective, cross-sectional database study. Patients were assigned to either the HCV or non-HCV cohort as described in the main text (see also Figure S1). The diagnosis and product codes used to define HCV are listed in Tables S1 and S2. Cross-sections were extracted on a rolling basis with between January 2015 and February 2020 with the final cross-section designed to have a non-overlapping selection window to facilitate subsequent validation of the model, see Figure S2. The ML algorithm is depicted above for a single cross-section. It shows medical history "seen" by the algorithm in the 24 month look back for the patient and how the algorithm predicts a HCV diagnosis in the 12 month selection window.

Figure S 2 Rolling cross-sectional study design.



1.1 Machine Learning Algorithm: Implementation and Validation

The GBT algorithm was executed using the XGBoost (xgboost v1.2.1) implementation for Python (3.6.8). The GBT algorithm was trained using cross-sections 1 to 14 and subsequently tested on cross-section 26, where the selection window did not overlap with the training cross-section selection windows.

The non-HCV cohort was randomly down-sampled to a ratio of 100 non-HCV to HCV patients within each cross-section. This ratio was chosen to reduce the class imbalance whilst preserving the heterogeneity of the non-HCV cohort. After down-sampling, a selection criterion that requires each patient to have a predictor in the lookback period was applied. When assessing the model performance on the test cross-section, the number of non-HCV patients was rescaled to the ratio seen within the underlying population. This was to account for the artificially low number of non-HCV patients which would result in an artificially low false positive rate.

Model complexity was optimised by reducing the number of features iteratively. An initial model was trained on the full predictor space (931 predictors) using the earliest two cross-sections. This model was applied to cross-section 14, a left out and non-overlapping training cross-section, and performance was reported as improvement in precision over Universal Screening at recall levels of 5%, 10%, 20%, 50% and 75%. Subsequent models were retrained iteratively, reducing the predictor space to only the most important predictors as identified by the total gain, i.e. the contribution of splitting on the predictor to model performance. The model with the lowest number of predictors without any reduction in performance was chosen.

The training of GBT algorithm included hyperparameter tuning for the learning rate, the number of estimators, max depth, min child weight, and gamma using the grid search method in a cross-validated manner.

The final step involved training the GBT algorithm with the optimised hyperparameters on all available training data followed by its application to the held-out cross-section to assess model performance.

2 Supplementary Results

Figure S3 shows model performance as improvement over Universal Screening versus model complexity (the number of model features) at recall levels of 5%, 10%, 20%, 50% and 75%. The 100-predictor model was chosen as it reduced complexity whilst retaining model performance.

Figure S 3 Performance versus complexity (number of predictors).

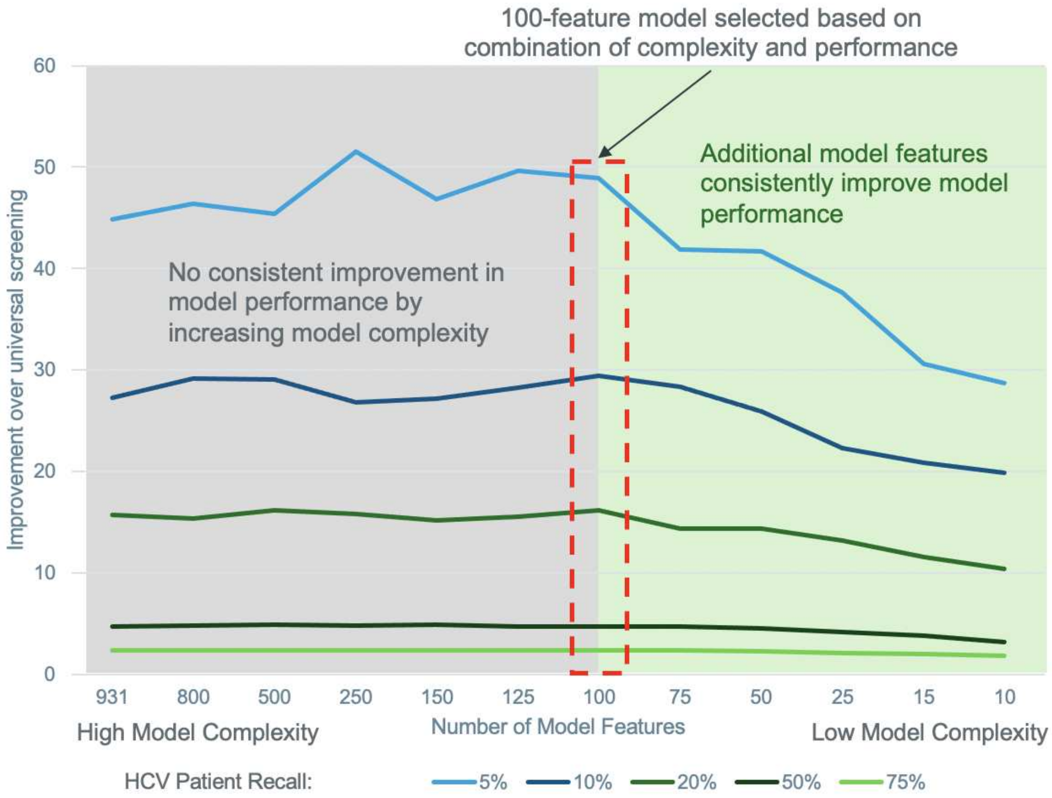


Figure S 4 Contribution of age and gender to the HCV risk score where each patient is represented to a single data point.

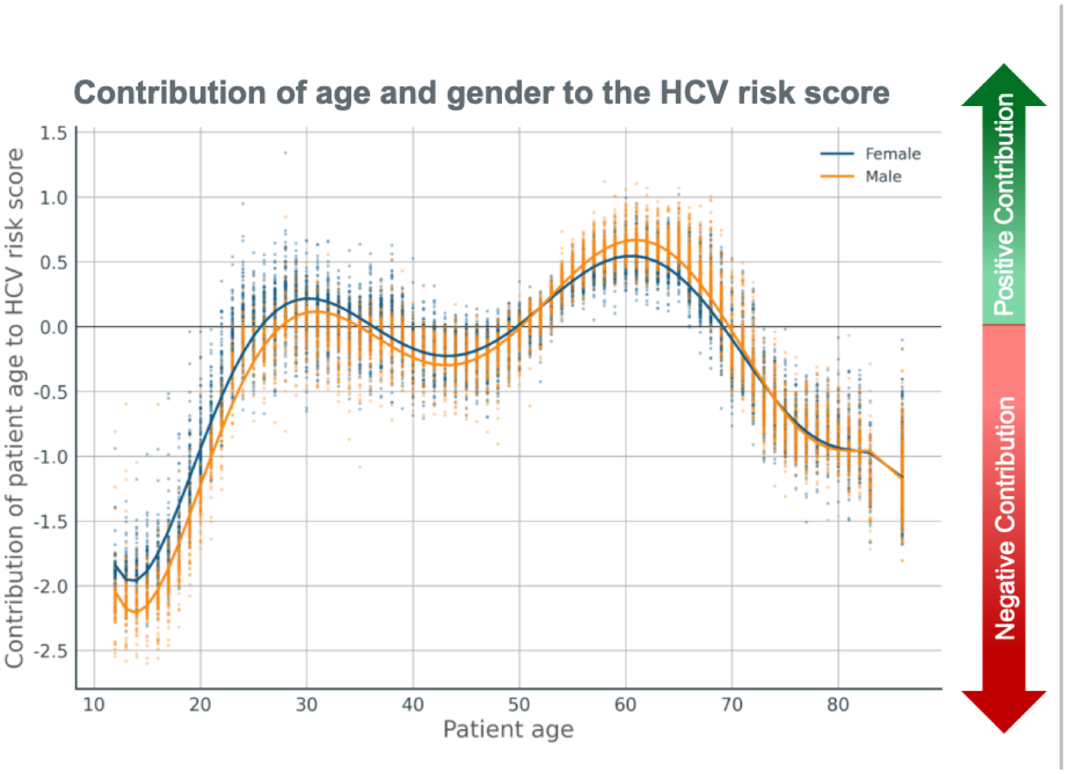


Figure S 5 False Negative Rate per HCV Patient Recall post correction for bias in Age by the protected characteristics; a) Age b) Gender and c) Race

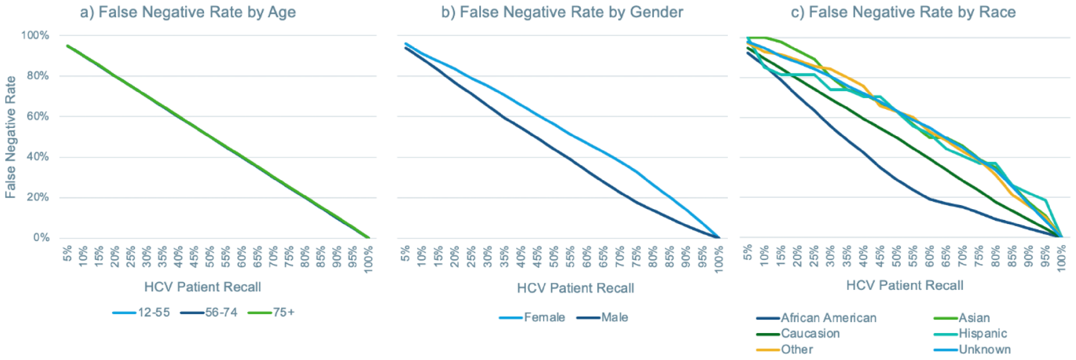


Figure S 6 False Negative Rate per HCV Patient Recall post correction for bias in Gender by the protected characteristics; a) Age b) Gender and c) Race

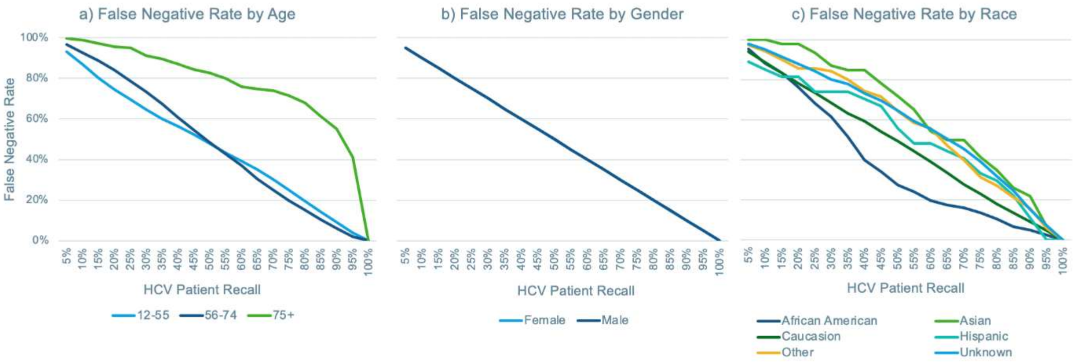
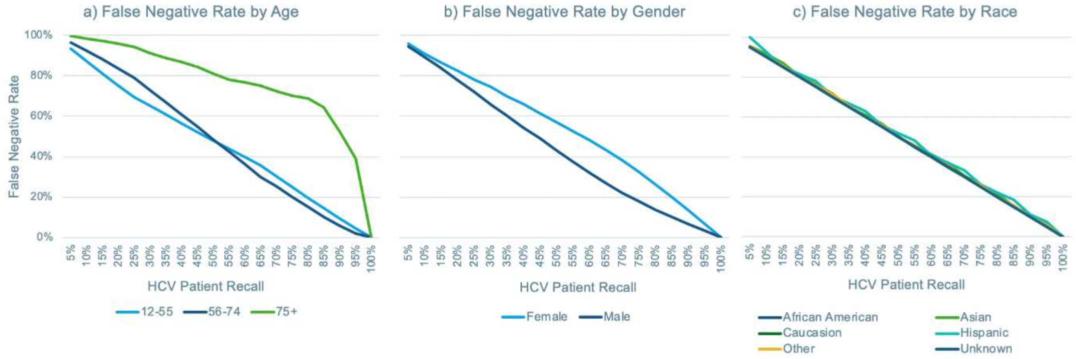


Figure S 7 False Negative Rate per HCV Patient Recall post correction for bias in Race by the protected characteristics; a) Age b) Gender and c) Race



3 Supplementary Tables

3.1 Diagnosis codes and prescription products for HCV

Table S 1 List of ICD 9 and ICD 10 codes used to select HCV patients.

DIAGNOSIS CODE TYPE	DIAGNOSIS CODE	DIAG DESCRIPTION
ICD 9	070.41	ACUTE HEPATITIS C WITH HEPATIC COMA
ICD 9	070.44	CHRONIC HEPATITIS C WITH HEPATIC COMA
ICD 9	070.51	ACUTE HEPATITIS C WITHOUT MENTION OF HEPATIC COMA
ICD 9	070.54	CHRONIC HEPATITIS C WITHOUT MENTION OF HEPATIC COMA
ICD 9	070.7	UNSPECIFIED VIRAL HEPATITIS C
ICD 9	070.70	UNSPECIFIED VIRAL HEPATITIS C WITHOUT HEPATIC COMA
ICD 9	070.71	UNSPECIFIED VIRAL HEPATITIS C WITH HEPATIC COMA
ICD 9	V02.62	CARRIER OR SUSPECTED CARRIER OF HEPATITIS C
ICD 10	B17.1	ACUTE HEPATITIS C
ICD 10	B17.10	ACUTE HEPATITIS C WITHOUT HEPATIC COMA
ICD 10	B17.11	ACUTE HEPATITIS C WITH HEPATIC COMA
ICD 10	B18.2	CHRONIC VIRAL HEPATITIS C
ICD 10	B19.2	UNSPECIFIED VIRAL HEPATITIS C
ICD 10	B19.20	UNSPECIFIED VIRAL HEPATITIS C WITHOUT HEPATIC COMA
ICD 10	B19.21	UNSPECIFIED VIRAL HEPATITIS C WITH HEPATIC COMA
ICD 10	Z22.52	CARRIER OF VIRAL HEPATITIS C

Table S1 List of products used to define treatment for HCV.

Generic Product ID (10) DESCRIPTION	MARKETED PRODUCT NAME
BOCEPREVIR	VICTRELIS
DACLATASVIR DIHYDROCHLORIDE	DAKLINZA
ELBASVIR-GRAZOPREVIR	ZEPATIER
GLECAPREVIR-PIBRENTASVIR	MAVYRET
INTERFERON ALFA-2B	INTRON A
	INTRON A W/DILUENT
INTERFERON ALFACON-1	INFERGEN
LEDIPASVIR-SOFOSBUVIR	HARVONI
	LEDIPASVIR/SOFOSBUVIR
OMBITASVIR-PARITAPREVIR-RITONAVIR	TECHNIVIE
OMBITASVIR-PARITAPREVIR-RITONAVIR-DASABUVIR	VIEKIRA PAK
	VIEKIRA XR
PEGINTERFERON ALFA-2A	PEGASYS
	PEGASYS PROCLICK
PEGINTERFERON ALFA-2B	PEG-INTRON
	PEG-INTRON REDIPEN
	PEG-INTRON REDIPEN PAK 4
	PEGINTRON
RIBAVIRIN (HEPATITIS C)	COPEGUS
	MODERIBA

	MODERIBA 1200 DOSE PACK
	MODERIBA 800 DOSE PACK
	REBETOL
	RIBASPHERE
	RIBASPHERE RIBAPAK
	RIBATAB
	RIBAVIRIN
SIMEPREVIR SODIUM	OLYSIO
SOFOSBUVIR	SOVALDI
SOFOSBUVIR-VELPATASVIR	EPCLUSA
	SOFOSBUVIR/VELPATASVIR
SOFOSBUVIR-VELPATASVIR-VOXILAPREVIR	VOSEVI
TELAPREVIR	INCIVEK

3.2 List of predictors concepts

Table S2 List of predictors used for creating features for the ML algorithm.

PREDICTOR CONCEPTS
AGE
GENDER
RACE
ABDOMINAL CT SCAN
ABDOMINAL SURGERIES

ABNORMAL STOOL COLOR
ABNORMAL WGT LOSS
ACL INHIBITORS
ADDICTION MEDICINE SPECIALTY VISIT
ALCOHOL USE ABUSE DEPENDENCE
ALCOHOL WITHDRAWAL
ALCOHOLIC LIVER DISEASE
ALOPECIA AREATA
ALPHA 1 ANTITRYPSIN DEFICIENCY
AMBULATORY SPECIALTY VISIT
AMEBICIDES
AMINOGLYCOSIDES
ANALGESICS
ANOREXIA
ANTHELMINTICS
ANTI INFECTIVE AGENTS
ANTI INFLAMMATORY ANALGESICS
ANTI MOTILITY DRUGS
ANTI REJECTION AGENTS
ANTIANXIETY AGENTS
ANTIDEPRESSANTS
ANTIDIARRHEAL PROBIOTIC AGENTS

ANTIEMETICS
ANTIFUNGALS
ANTIHYPERLIPIDEMICS COMBOS
ANTIHYPERLIPIDEMICS MISC
ANTIMALARIALS
ANTIMYOBACTERIAL AGENTS
ANTIPHOSPHOLIPID SYNDROME
ANTIPSYCHOTICS ANTIMANIC AGENTS
ANTIRETROVIRALS
ANTIULCERANTS
ANTIULCERANTS PPIS
ANXIETY
ARTERITITS
ARTHROPOD BORNE HEMMORRHAGIC FEVER
ASCITES
AUTOIMMUNE HEMOLYTIC ANEMIA
B-CELL NON HODGKINS LYMPHOMA
BACTEREMIA
BARIATRIC SPECIALTY VISIT
BEHAVIORAL HEALTH SPECIALTY VISIT
BENIGN NEOPLASM
BILE ACID SEQUESTRANTS

BMT SCT TRANSPLANT
BRUISING
CACHEXIA
CARDIOPULMONARY BYPASS
CELIAC DISEASE
CEPHALOSPORINS
CHEST PAIN
CHLAMYDIA
CHOLANGITIS
CHOLESTEROL ABSORPTION INHIBITORS
CHOLESTEROL AGENTS
CHRONIC FATIGUE
CHRONIC LIVER DISEASE
CHRONIC LUNG DISEASE
CHURG STRAUSS SYNDROME
CIRRHOSIS
CKD ESRD
CLINICAL SOCIAL WORKER SPECIALTY VISIT
COLITIS
COLON CANCER SCREENING
COLONOSCOPY
COMPLETE BLOOD COUNT

CONFUSION
CONVULSIONS
COUNSELOR SPECIALTY VISIT
CRITICAL CARE SPECIALTY VISIT
CRYOGLOBULINEMIA
CYTOMEGALOVIRUS
DARK URINE
DEPRESSION
DIABETES
DIAGNOSTIC TESTING SPECIALTY VISIT
DIARRHEA
DIURETICS
DROWSINESS
DRUG SUBSTANCE WITHDRAWAL
DRY EYES
DYSARTHRIA
DYSMENORRHEA
DYSPEPSIA
DYSPNEA
EARLY SATIETY
EDEMA
EMERGENCY MEDICINE SPECIALTY VISIT

ENTERITIS DUE TO UNSPECIFIED VIRUS
EPIDEMIOLOGY PUBLIC HEALTH SPECIALTY VISIT
ERYTHROPOIESIS STIMULATING AGENTS ESAS
FAMILIAL HCV
FAMILY PRACTICE SPECIALTY VISIT
FEVER
FIBRATES
FIBROMYALGIA
FLU VACCINES
FLUOROQUINOLONES
GASTROENTEROLOGY SPECIALTY VISIT
GENERAL PRACTICE SPECIALTY VISIT
GENERAL SURGERY SPECIALTY VISIT
GENETICS SPECIALTY VISIT
GERD
GERIATRIC MEDICINE SPECIALTY VISIT
GLOMERULONEPHRITIS
GONORRHOEAE
GRANULOMATOSIS WITH POLYANGIITS
HCV TESTS
HEADACHE
HEART PALPITATIONS

HEARTBURN
HEMATOLOGY SPECIALTY VISIT
HEMATURIA PROTEINURIA URINALYSIS
HEMOCHROMATOSIS
HEMODIALYSIS
HEMODIALYSIS TREATMENT
HEMOPHILIA
HEMORRHOIDS
HEPATIC CARCINOMA
HEPATIC ENCEPHALOPATHY
HEPATIC FIBROSIS
HEPATIC OSTEODYSTROPHY
HEPATITIS CO INFECTION
HEPATITIS VACCINES
HEPATOLOGY SPECIALTY VISIT
HEPATOMEGALY SPLENOMEGALY
HERPES SIMPLEX VIRUS
HIGH RISK SEXUAL BEHAVIOR
HISTORY OF CARDIOPULMONARY BYPASS CABG
HIV AIDS
HOMELESSNESS ECONOMIC BURDEN
HUMAN PAPILLOMAVIRUS

HYPERGLYCEMIA
HYPERLIPIDEMIA
HYPERTENSION
HYPERTHYROIDISM
HYPOGLYCEMIA
HYPOTHYROIDISM
IBS
IMMUNE THROMBOCYTOPENIC PURPURA
IMMUNOLOGY SPECIALTY VISIT
IMMUNOSUPRESSIVES
INCARCERATION HISTORY
INFECTIOUS DISEASE SPECIALTY VISIT
INFECTIOUS MONONUCLEOSIS
INFLUENZA
INJECTABLE IRON
INSOMNIA
INSULIN RESISTANCE
INTERNAL MEDICINE SPECIALTY VISIT
IPF
JAUNDICE
JOINT PAIN
JUGULAR VEIN DISTENTION

LAB RESULT – ALT
LAB RESULT – AST
LAB RESULT – BILIRUBIN
LACTOSE INTOLERANCE
LICHEN PLANUS
LIVER ABSCESS
LIVER BIOPSY
LIVER DISEASE MULTIANALYTE ASSAYS
LIVER ELASTOGRAPHY
LIVER FAILURE
LIVER FUNCTION STUDIES
LOWER RESP TRACT INFECTION
LUPUS
LYMPHADENOPATHY
MACROLIDES
MAMMOGRAPHY
MICROSCOPIC POLYANGIITS
MILITARY SERVICE
MISC ANTI-INFECTIVE AGENTS
MOORENS CORNEAL ULCERS
MOSQUITO BORNE VIRAL ENCEPHALITIS
MTP INHIBITORS

MYALGIAS
NAUSEA VOMITING
NECROLYTIC ACRAL ERYTHEMA
NEUROPATHY
NICOTINIC ACID DERIVATIVES
NON-ALCOHOLIC STEATOHEPATITIS NASH
NON-HODGKIN LYMPHOMA
NON-INFECTIOUS HEPATITIS
NON0NARCOTIC ANALGESICS
NURSE PRACTITIONER SPECIALTY VISIT
OBESITY
OBSTETRICS GYNECOLOGY SPECIALTY VISIT
OCCUPATIONAL EXPOSURE
OPIOID ANALGESICS
ORGAN TRANSPLANT
OSTEOARTHRITIS
OTHER ABDOMINAL PAIN
OTHER ANTIVIRALS
OTHER DRUG USE ABUSE
OTHER FATIGUE
OTHER HEADACHE SYNDROMES
OTHER HEMORRHAGIC CONDITIONS

OTHER MALAISE
OTHER POXVIRUS INFECTIONS
OTHER PURPURA
OTHER STI
OTHER VASCULITIS
PCSK9 INHIBITORS
PENICILLINS
PEPTIC ULCER DISEASE
PHYSICIAN ASSISTANT SPECIALTY VISIT
POLYMYOSITIS DERMATOMYOSITIS
PORPHYRIA CUTANEA TARDA
PORTAL HYPERTENSION
PPIS
PROSTATE CANCER
PRURITUS
PSORIASIS
PSYCHIATRY SPECIALTY VISIT
PULMONOLOGY SPECIALTY VISIT
RAPE
RASH
RAYNAUDS PHENOMENON
REACTIVE ARTHRITIS

REGISTERED NURSE SPECIALTY VISIT
RENAL CANCER
RHEUMATOID ARTHRITIS
RHEUMATOID VASCULITIS
RIGHT SIDED HF
RIGHT UPPER ABDOMINAL PAIN
RISK OF INTRAVENOUS DRUG USE ABUSE
SCLERITIS
SCLERODERMA
SEDATIVES HYPNOTICS SLEEP DISORDER AGENTS
SENSORY NEUROPATHY
SJOGRENS DISEASE
SKIN ABCESS
SLOW VIRUS INFECTIONS
SLURRED SPEECH
SPIDER ANGIOMAS NEVUS
SPLENOMEGALY
STATINS
STD TESTS
STEATORRHEA
STEATOSIS
STREPTOCOCCUS PNEUMONIAE

SUBSTANCE USE ABUSE DEPENDENCE
SUBSTANCE USE DISORDER AGENTS
SULFONAMIDES
SWELLING OF LIMB
SYPHILIS
TETRACYCLINES
THALASSEMIA
THROMBOCYTOPENIA
THROMBOSIS
THYROIDITIS
TRANSEXUALISM
TRANSFUSIONS
TRANSVESTIC FETISHISM
TRICHOMONIASIS
ULCER THERAPY COMBOS
UNDERWEIGHT
UPPER ENDOSCOPY
UPPER RESPIRATORY TRACT INFECTION
URINARY RETENTION
UVEITIS
VACCINES (HEPATITIS or INFLUENZA)
VARICES

VIRAL CHLAMYDIAL INFECTIONS
VIRAL HEPATITIS
VIRAL PNEUMONIA
VITAMIN D DEFICIENCY
VITILIGO
WEAKNESS
XANTHELASMA XANTHOMA