

Ontology Development and Use for Cholangiocarcinoma Risk Factors and Predictions : A Term Enrichment Data Analysis and Machine Learning Classification

Anuwat Pengput, Alexander Diehl

Submitted to: JMIR Cancer
on: April 23, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Ontology Development and Use for Cholangiocarcinoma Risk Factors and Predictions : A Term Enrichment Data Analysis and Machine Learning Classification

Anuwat Pengput^{1,2} MPH, MSc, PhD; Alexander Diehl¹ PhD

¹Biomedical Informatics Jacobs School of Medicine and Biomedical Sciences University at Buffalo, State University of New York Buffalo US

²Department of Community Public Health Faculty of Public Health and Allied Health Science Sirindhorn College of Public Health Khon Khan Khon Khan TH

Corresponding Author:

Alexander Diehl PhD
Biomedical Informatics
Jacobs School of Medicine and Biomedical Sciences
University at Buffalo, State University of New York
77 Goodell Street
Buffalo
US

Abstract

Background: Cholangiocarcinoma (CCA) is a critical public health problem in Thailand. The prevalence is much higher than other areas in the world. Data about CCA are stored in different data sources and standards between research-based data collection and electronic health records (EHR).

Objective: This study aims to integrate and analyze cholangiocarcinoma (CCA) data from various sources to investigate risk factors and develop prediction models using the Cholangiocarcinoma Ontology (CCAO).

Methods: Datasets from Thailand were annotated with CCAO and analyzed using ontology-based term enrichment methods. We apply ontology term enrichment analysis, similar to that used with the Gene Ontology, for identifying significant risk factors for suspected CCA and patients with CCA. Our program provides a list of significant terms associated with CCA and a visualization of ontology hierarchy with significant terms highlighted. The outputs of the term enrichment analyses have been used as the inputs to machine learning classification tasks.

Results: The results confirmed that indicators for CCA include dilated bile ducts, periductal fibrosis, and hepatic mass, based on ultrasound findings from several years prior. Our analysis also revealed demographic and lifestyle risk factors such as male gender, having no education, alcohol consumption, smoking, being a farmer, and having diabetes. We seeded a random forest classifier with the term enrichment results and predicted CCA patients with average 0.92 precision-recall curve score (0.023 standard deviation) with age, dilated bile ducts, periductal fibrosis, suspected CCA, and hepatic mass as the top five important features.

Conclusions: These findings can be used to focus and monitor populations at risk for CCA. Expanding CCAO with molecular data related to CCA using ontology-driven term enrichment analysis and machine learning will help us to discover new hypotheses to decrease the morbidity and mortality of CCA in Thailand.

(JMIR Preprints 23/04/2025:76443)

DOI: <https://doi.org/10.2196/preprints.76443>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to the public.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/76443>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/76443>



Original Manuscript

Original Paper

Anuwat Pengput ^{1,2}, MPH, MSc, PhD; Alexander D. Diehl ¹, PhD

1 Department of Biomedical Informatics, University at Buffalo, Buffalo, New York, USA

2 Department of Community Public Health, Sirindhorn College of Public Health Khon Kaen, Faculty of Public Health and Allied Health Science, Praboromarajchanok Institute, Khon Kaen, Thailand

Corresponding Author:

Alexander D. Diehl, PhD

Department of Biomedical Informatics,
Jacobs School of Medicine and Biomedical Sciences,
University at Buffalo

77 Goodell Street

Buffalo, New York 14203

Phone: 716-881-8975

addiehl@buffalo.edu

Ontology Development and Use for Cholangiocarcinoma Risk Factors and Predictions : A Term Enrichment Data Analysis and Machine Learning Classification

Abstract

Background: Cholangiocarcinoma (CCA) is a critical public health problem in Thailand. The prevalence is much higher than other areas in the world. Data about CCA are stored in different data sources and standards between research-based data collection and electronic health records (EHR).

Objective: This study aims to integrate and analyze cholangiocarcinoma (CCA) data from various sources to investigate risk factors and develop prediction models using the Cholangiocarcinoma Ontology (CCAO).

Methods: Datasets from Thailand were annotated with CCAO and analyzed using ontology-based term enrichment methods. We apply ontology term enrichment analysis, similar to that used with the Gene Ontology, for identifying significant risk factors for suspected CCA and patients with CCA. Our program provides a list of significant terms associated with CCA and a visualization of ontology hierarchy with significant terms highlighted. The outputs of the term enrichment analyses have been used as the inputs to machine learning classification tasks.

Results: The results confirmed that indicators for CCA include dilated bile ducts, periductal fibrosis, and hepatic mass, based on ultrasound findings from several years prior. Our analysis also revealed demographic and lifestyle risk factors such as male gender, having no education, alcohol consumption, smoking, being a farmer, and having diabetes. We seeded a random forest classifier with the term enrichment results and predicted CCA patients with average 0.92 precision-recall curve score (0.023 standard deviation) with age, dilated bile ducts, periductal fibrosis, suspected CCA, and hepatic mass as the top five important features.

Conclusions: These findings can be used to focus and monitor populations at risk for CCA.

Expanding CCAO with molecular data related to CCA using ontology-driven term enrichment

analysis and machine learning will help us to discover new hypotheses to decrease the morbidity and mortality of CCA in Thailand.

Keywords: Biomedical Ontology, Ontology, Term Enrichment Analysis, Machine Learning, Cholangiocarcinoma

Introduction

Cholangiocarcinoma (CCA) is a major public problem in Thailand. The prevalence of CCA in Thailand is much higher than in other areas worldwide, especially in the northeastern Thailand. The main risk factor for CCA is liver fluke, infections, which is associated with eating raw, fermented, pickled, and undercooked cyprinid fish, which is a cultural tradition in the region. *O. viverrini* infections produce hepatic bile duct and portal connective tissue inflammation. Chronic infections and inflammation have been indicated to be risk factors for carcinogenesis [1, 2].

Data about *O. viverrini* and CCA (OVCCA) are deposited in several repositories in Thailand. The Cholangiocarcinoma Screening and Care Program (CASCAP) maintains one of the largest and most comprehensive databases on OVCCA. Data is collected based on six data collection forms: CCA-01, "Demographic Information, Enrollment," CCA-02, "Ultrasound," CCA-02.1, "Confirmatory Diagnosis," CCA-03, "Diagnosis and Treatment," CCA-04, "Final Staging Diagnosis," and CCA-05, "Post Operation Follow-up" [3, 4]. Moreover, electronic health records (EHR) from general and community hospitals in Thailand also contain data and information on patients with or suspected of suffering from CCA including symptoms, clinical findings, treatments, and diagnoses.

These databases store data elements using different standards. For EHRs, the International Statistical Classification of Disease and Related Health Problems, 10th revision, Thai Modification (ICD-10-TM) is being used for morbidity and mortality coding in health services statistics, as well as for billing and payment [5, 6]. In contrast, the CASCAP study collects data and information as research-based data elements that capture details about CCA more specific than those in ICD-10-TM. Working with data from different sources and standards poses challenges of comparability.

In order to understand the complexity of risk factors for CCA and to integrate data from different sources, the Cholangiocarcinoma Ontology (CCAO) [7] has been built as an application ontology under the Basic Formal Ontology (BFO) [8]. CCAO relies upon BFO to provide an upper-level framework to structure the ontology. CCAO represents all data elements in the research cohort and electronic health records (EHR) from the Kalasin Provincial Public Health Office, Thailand, and contains more than 430 classes specific to the CCA, and also includes classes from external ontologies such as the Ontology for General Medical Sciences (OGMS) [9], and the Uber-anatomy ontology (Uberon) [10].

Ontologies have long been used to designate all entities within an area of reality and all relationships between those entities in a way that makes them interpretable by both humans and computers [11]. One of the main uses of ontology is to conduct term enrichment analysis (TEA) on sets of genes. TEA is commonly used with the Gene Ontology (GO) to determine statistically significant over- or under-representation of GO terms associated with a particular set of genes [12], to identify particular biological processes, molecular functions, or cellular components that may be of importance in a particular experimental setting. If a certain set of genes is up-regulated under certain conditions, the enrichment analysis allows for the development of new hypotheses regarding the experimental system that can be tested through further experimentation [13]. Studying gene-disease and disease-disease associations provides insights into analysis for the high-throughput data such as a list of genes or genes products to elucidate molecular processes of complex diseases, help to describe biological questions in the disease context, and discover unanticipated functions [14].

TEA has three categories: first-generation - over-representation analysis (ORA), second-

generation - Functional Class Scoring (FCS), and third generation - Pathway-Topology Based (PBD). This study focuses on the ORA approach, which we use to develop our computing program and conduct data analysis. The ORA provides the results with a list of the most relevant ontology terms, labeled as over- or under-represented, and then ordered based on multiple-test corrected *P*-values. The main advantage of using ORA is its ability to generate hypotheses which can be tested them using other statistical or experimental methods. ORA remains one of the most common methods for hypothesis generation from high-throughput datasets [15, 16]. ORA can perform a hypothesis-driven task, such as asking whether an association with a particular disease is significantly enriched or depleted in a set of genes.

There are a few tools outside GO that develop and apply TEA methods to profile other sets of interest, such as patient cohorts from electronic medical records, using a variety of ontologies including SNOMED CT, Ontology for General Medical Sciences (OGMS), and other biomedical ontologies [12]. TEA [12, 17, 18] is not only a useful tool for analyzing the results of high-throughput experiments, but also TEA can generate hypotheses for clinical and public health research. TEA can be utilized as an exploratory tool to generate hypotheses for clinical research by analyzing data annotations with various biomedical ontologies on medical records or biomedical research databases in conjunction with data annotations of molecular datasets. Using TEA to identify common pairs of ontology terms can help to find combinations of demographic information and comorbidities, or risk factors and clinical findings that are in common in the population.

Our work aims to develop and apply the TEA workflow implementing to clinical, and public health datasets by integrating with well-defined sources of biomedical ontologies. Moreover, we believe the ontology-driven ORA method can facilitate and improve the model performances of machine learning classifications for participants with suspected cholangiocarcinoma (CCA) and patients with CCA.

Methods

Development of Cholangiocarcinoma Ontology

The Cholangiocarcinoma Ontology (CCAO) has been developed in OWL format using Protégé based on data about demographics, ultrasound findings, diagnoses and treatments, and post-operative outcomes from areas of Thailand where CCA is endemic due to *O. viverrini* [7]. CCAO also represents CCA diagnosis data from EHRs (Figure 1). Datasets from Thailand were annotated with CCAO and analyzed using ontology term enrichment to compare specific subpopulations to the participant group as a whole.

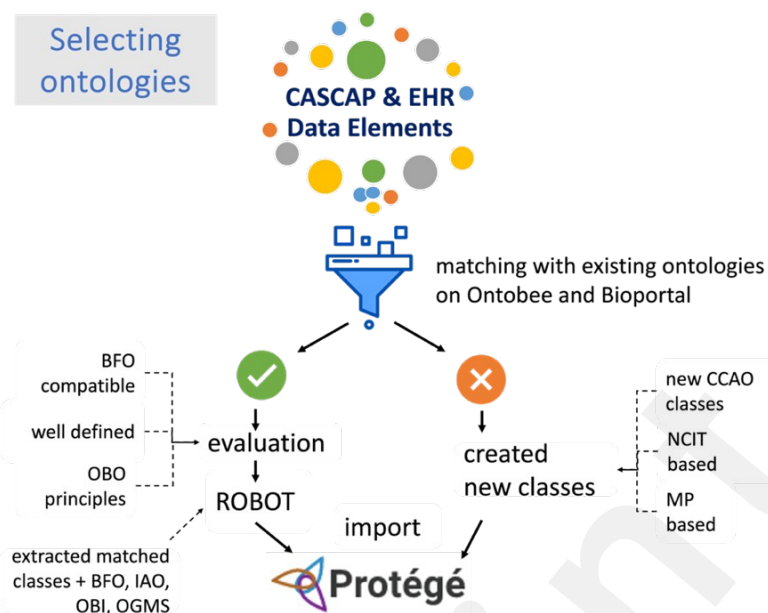


Figure 1. Flow diagram of selecting ontologies for CCAO.

Sample

We received IRB approval (UB IRB ID: STUDY00006059) from the University at Buffalo to obtain and de-identify data from the Kalasin Provincial Public Health Office (KPH). We also received approval from the Chief Executive (Physician) of KPH to obtain and use the data in our study. The data included participants who live in Kalasin, Thailand, and who participated in verbal and ultrasound (VUS) screenings with KPH. Also via KPH, we received data from Electronic Health Records (EHR) databases of individuals who were diagnosed with CCA, and other disorders and symptoms related to CCA. Data cover the study period between January 1, 2016, and May 31, 2022.

Data management

We use R version 4.2.2 with package ‘Tidyverse’ within RStudio for data wrangling and management. We imported the verbal screening dataset containing 56,805 observations and the ultrasound screening dataset containing 35,570 observations. Per our approved IRB, we performed exact matching using citizen identification numbers (CID) to identify patients across verbal screening, ultrasound screening databases and EHR databases. We deidentified the identifiable data to coded data based on the IRB approval using R package ‘anonymizer’ version 0.2.2 [19]. We merged verbal and ultrasound screening datasets using CID to identify participants who participated in both screenings. We excluded the incomplete records and outlier participants under 20 years old and over 100 years old. The final datasets contain 21,038 participants receiving VUS screenings and 676 patients visiting hospitals with participated in VUS screenings. Per the IRB, the original data was secured and not referenced again during the course of the study.

Creating ontology classes to represent EHR data elements

Data elements from the EHR dataset contain diagnoses related to CCA of patients who visited the hospitals during the study period. The diagnoses are represented by International Statistical Classification of Diseases and Related Health, 10th. Revision Thai Modification (ICD-10 TM) in the EHR databases with twenty-three ICD-10 TM codes used. We created new CCAO terms under ‘OGMS: Clinical finding’ to express the data elements of the ICD-10 TM codes and labels. For instance, we created a class ‘CCAO_00436: finding about intrahepatic cholangiocarcinoma’ with a logical definition ‘*finding about cholangiocarcinoma*’ and ‘*(is about) some ‘intrahepatic*

cholangiocarcinoma') to represent 'ICD10-TM_C221: Intrahepatic bile duct carcinoma, Cholangiocarcinoma'. All new terms were imported to CCAO using Protégé [20].

Data annotation

We developed a Python script to generate data subsets as inputs for term enrichment analysis using Python 3.9.12. The script annotates data values of variables in the prepared datasets to the ontology classes in CCAO. For instance, participants with suspected CCA (value = 1) were annotated with 'CCAO_00292: finding about suspected cholangiocarcinoma'. The script can generate either a code or label output version based on the selected variables and values as a text file. For inputs of term enrichment analysis, we generate two text files consisting of a universal set, which includes data elements of all participants annotated to CCAO classes, and an experimental set, which includes subgroups of interest among all participants annotated to CCAO classes.

Statistical Analysis

Descriptive analysis

Descriptive statistics were used to describe demographic characteristics of participants. We used frequency and percentage for categorical data. For continuous data, we used mean and standard deviation (SD) if data was normal distribution and used median and min-max if data was non-normal distribution.

Term enrichment analysis

Data annotated to CCAO classes were analyzed using TEA to compare the specific subpopulations that match certain criteria, such as age group, gender, or OV infection status, to the participant group as a whole. For TEA, *P*-values were calculated using the hypergeometric distribution test and were adjusted using the Bonferroni correction for multiple testing [16]. The results of the program were validated with MATLAB, Microsoft Excel, and other online statistics calculators. The program for TEA was developed previously by Alexander Diehl using the Perl programming language [21]. The program creates the output file of terms with summed up annotations and *P*-values. The output files are used to generate a visualized hierarchy of the result using Graphviz version 2.36 [22]. Significant *P*-values less than 0.05 are represented as over-represented and under-represented classes using different shades of colors with correspond to the magnitude of the *P*-values ().

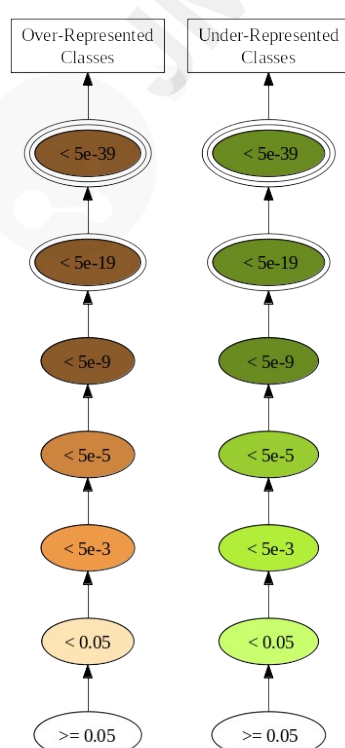


Figure 2. *P*-values Color Legend for *P*-values of term enrichment results.

Machine Learning Classifications

We performed the machine learning classifications for participants with suspected CCA and patients with CCA using the PyCaret library version 3.1.0 in Python [23]. PyCaret is an open source, low-code machine learning library in Python. The library is an end-to-end machine learning and model management tool and a Python wrapper around several machine libraries and framework such as scikit-learn, XGBoost, LightGMB, CatBoost, spaCy, and etc.

For suspected CCA classification, we included 21,038 participants who participated in verbal and ultrasound screening in the data analyses. The dataset contained 100 variables which included 99 features and designated suspected CCA as the target variable. The target variable is binary data in which '0' is non-suspected CCA and '1' is suspected CCA. Features were selected based on statistically significant over-represented classes from the term enrichment results for suspected CCA plus age (Figure 3), resulting in the inclusion of 45 of 101 features for developing classification models (Indicator model). In order to control for the effect of main indicator variables used in the diagnoses of suspected CCA, we excluded dilated bile duct, periductal fibrosis, and hepatic mass variables from the models (Non-indicator model) and compared the model performance between the indicator models and non-indicator models using Student's t-test with a significance level of 0.05.

We used the "*ClassificationExperiment()*" function from PyCaret and split data 70% for training set ($n = 14,726$), and 30% for test set ($n = 6,312$), respectively. We trained and evaluated parameters using stratified 10-fold cross validation. The various classifiers that were used in the comparison models were the gradient boosting classifier, logistic regression, random forest classifier, light gradient boosting machine, linear discriminant analysis, and naive Bayes. Based on 10 separate executions of the classifier using different random number seeds, mean and standards deviation of AUROC, AUPRC, precision, recall, and F1 were calculated for model selection and evaluation in order to control for effect of using different random number seeds in the machine learning models [24, 25].

For CCA classification in subjects who went on to develop CCA, we included 676 patients in the data analyses who participated in verbal and ultrasound screening and who visited the hospitals and received a CCA related diagnosis as determined by assigned ICD codes. The dataset contained 112 columns which included 111 features and CCA as a target variable. The target variable is binary data which '0' is non-CCA and '1' is CCA. Features were selected based on statistically significant over-represented classes from the term enrichment results for CCA plus age at hospital visits (Figure 9). As a result, we included 40 features from original features for developing classification models (Indicators models). To control for the effect of main indicator variables used in the diagnoses of patients with CCA, we excluded dilated bile duct, periductal fibrosis, and hepatic mass, and suspected CCA variables from the models (Non-indicator model) and compared the model performance between the indicator models and non-indicator models using Student's t-test with a significance level of 0.05. We split the data 70% for training set ($n = 473$), and 30% for test set ($n = 203$), respectively. We then used the same procedure as the machine learning classification for suspected CCA.

Results

Baseline characteristics

The majority of participants in verbal and ultrasound screenings ($n = 21,038$) were female (60%) and 56 years old on average (Table 1). Most participants attended primary school (71%), and the main occupation was working as a farmer (81%). Almost a half of the participants reported having no fecal examination for liver flukes (48%), having had one fecal examination (34%), or having had more than one fecal examination (11%), respectively. More than half of the participants reported never received treatment with antiparasitic drug (64%), having once received treatment with antiparasitic drug (26%), or having more than once received treatment with antiparasitic drug (4.9%), respectively. Moreover, participants reported consumption of raw fresh-water fish or raw fermented fish (1.3%), smoking (22%), and alcohol drinking (38%).

Majority of participants in verbal screening, ultrasound screening, and visiting hospitals related to CCA ($n = 676$) were male (65%) and 64 years old on average (Table 2). Most participants attended primary school (79%), and the main occupation was working as a farmer (72%). More than half of the participants reported having no fecal examination for liver flukes (55%), having had one fecal examination (30%), or having had more than one fecal examination (8.4%), respectively. More than half of the participants reported never received treatment with antiparasitic drug (65%), having once received treatment with antiparasitic drug (25%), or having more than once received treatment with antiparasitic drug (4.9%), respectively. Moreover, participants reported consumption of raw fresh-water fish or raw fermented fish (5.6%), smoking (46%), and alcohol drinking (56%).

Table 1. Characteristics of participants in verbal and ultrasound screenings ($N = 21,038$)

Characteristic	Frequency (%)
Age	55.26 (9.88) ^a
< 40 years old	607 (3%)
40 – 50 years old	5,640 (27%)
51 – 60 years old	7,986 (38%)
> 60 years old	6,805 (32%)
Gender	
Male	8,350 (40%)
Female	12,688 (60%)
Highest education level	
No education	112 (0.5%)
Primary school	14,837 (71%)
Junior school	2,217 (11%)
Senior school	2,907 (14%)
Certificate degree	340 (1.6%)
Bachelor's degree	523 (2.5%)
Master or higher degree	102 (0.5%)
Occupation	
Unemployed	1,061 (5.0%)
Farmer	16,964 (81%)
Freelancer	1,491 (7.1%)
Private Business	605 (2.9%)
Civil Servant	550 (2.6%)
Other	367 (1.7%)
Fecal examination for liver flukes	

Characteristic	Frequency (%)
Never	10,144 (48%)
1 time	7,225 (34%)
2 times	2,347 (11%)
3 times	594 (2.8%)
More than 3 times	328 (1.6%)
I cannot remember	400 (1.9%)
Prior findings of liver fluke eggs	
Never examined	10,154 (48%)
Examined, but not found	5,677 (27%)
Examined and found	4,667 (22%)
I cannot remember	540 (2.6%)
Prior treatment with anti-parasite medicine (praziquantel)	
Never	13,540 (64%)
1 time	5,453 (26%)
2 times	1,032 (4.9%)
3 times	266 (1.3%)
More than 3 times	266 (1.3%)
I cannot remember	481 (2.3%)
Reported consumption of raw fresh-water fish or raw fermented fish	277 (1.3%)
Reported smoking	4,594 (22%)
Reported alcohol consumption	7,960 (38%)
^a Mean (SD)	

Table 2. Characteristics of participants participated in verbal and ultrasound screenings and visiting hospitals related to cca (N = 676)

Characteristic	Frequency (%)
Age	64.01 (12.47) ^a
< 40 years old	15 (2.2%)
40 – 50 years old	79 (12%)
51 – 60 years old	147 (22%)
> 60 years old	435 (64%)
Gender	
Male	438 (65%)
Female	238 (35%)
Highest education level	
No education	4 (0.6%)
Primary school	534 (79%)
Junior school	63 (9.3%)
Senior school	51 (7.5%)
Certificate degree	8 (1.2%)
Bachelor's degree	14 (2.1%)
Master or higher degree	2 (0.3%)
Occupation	
Unemployed	99 (15%)
Farmer	489 (72%)
Freelancer	45 (6.7%)
Private Business	17 (2.5%)
Civil Servant	20 (3.0%)
Other	6 (0.9%)
Fecal examination for liver flukes	
Never	375 (55%)
1 time	202 (30%)
2 times	57 (8.4%)
3 times	16 (2.4%)
More than 3 times	8 (1.2%)
I cannot remember	18 (2.7%)
Prior findings of liver fluke eggs	
Never examined	374 (55%)
Examined, but not found	148 (22%)
Examined and found	130 (19%)
I cannot remember	24 (3.6%)
Prior treatment with anti-parasite medicine (praziquantel)	
Never	440 (65%)
1 time	166 (25%)
2 times	33 (4.9%)
3 times	8 (1.2%)
More than 3 times	4 (0.6%)
I cannot remember	25 (3.7%)
Reported consumption of raw fresh-water fish or	38 (5.6%)

Characteristic	Frequency (%)
raw fermented fish	
Reported smoking	313 (46%)
Reported alcohol consumption	377 (56%)
^a Mean (SD)	

Term enrichment analyses

Suspected CCA group

For suspected CCA, 380 participants were annotated with 'finding about suspected cholangiocarcinoma' in CCAO based on ultrasound results among 21,038 participants in study group having both verbal screening and ultrasound results. In Figure 3, the TEA result indicated annotated classes were significantly over-represented among participants with suspected CCA including having alcoholism, alcohol consumption, smoking, no education, male gender, diagnosed with diabetes by clinician, diagnosed with other diseases by clinician, dilated bile duct, dilated common bile duct, dilated right lobe of bile duct, dilated left lobe of bile duct, cyst in kidney, cyst in left kidney, cyst in right kidney, left kidney parenchyma with atypical abnormal function, abnormal liver parenchymal echo, liver parenchyma with atypical abnormal structure, cirrhosis, periductal fibrosis type 1, periductal fibrosis type 2, periductal fibrosis type 3, hepatic cyst, hepatic mass low echo, left lobe hepatic mass low echo, right lobe hepatic mass low echo, hepatic mass high echo, left lobe hepatic mass high echo, right lobe hepatic mass high echo, hepatic mass mixed echo, left lobe hepatic mass mixed echo, right lobe hepatic mass mixed echo, gallbladder not seen, mass in wall of gallbladder, single mass in wall of gallbladder, thickening of wall of gallbladder, diffuse thickening of wall of gallbladder, stone in gallbladder, multiple stone in gall bladder, abdominal swelling, enlarged liver, appetite loss, weight loss, and jaundice (Figure 3 - Figure 8).

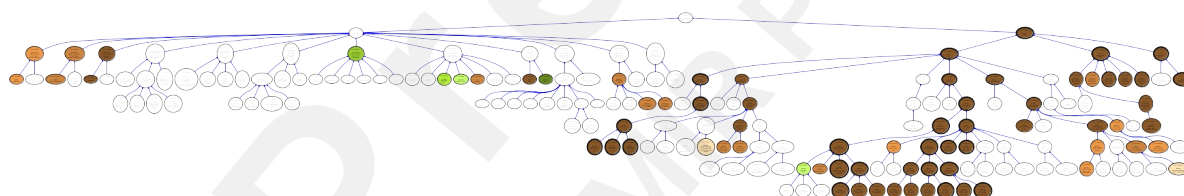


Figure 3. Term enrichment analysis results for the subgroup of having suspected CCA among participants who underwent verbal and ultrasound screening. A red arrow indicates the 'finding about suspected cholangiocarcinoma' class.

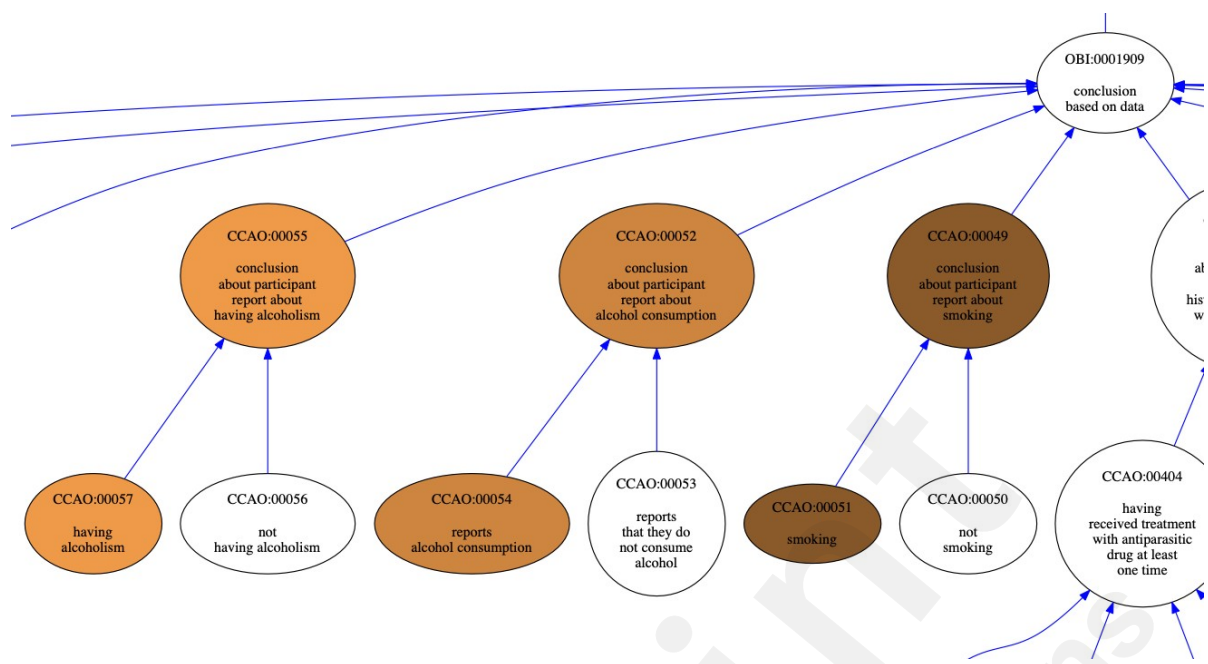


Figure 4. Zoomed-in term enrichment analysis results for the subgroup of having suspected CCA among participants who underwent verbal and ultrasound screening.

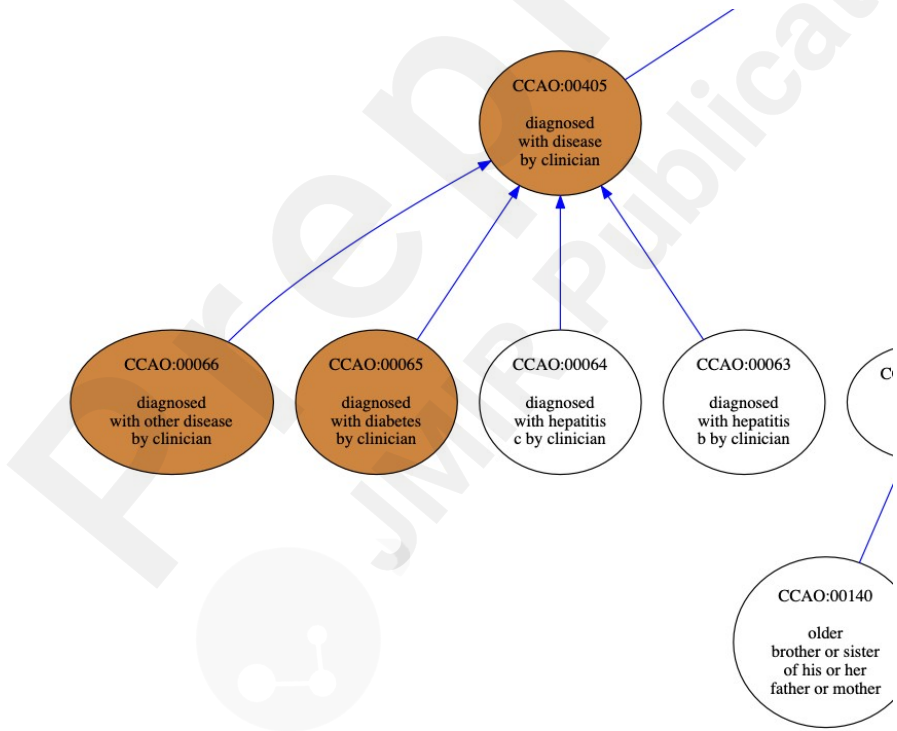


Figure 5. Zoomed-in term enrichment analysis results for the subgroup of having suspected CCA among participants who underwent verbal and ultrasound screening.

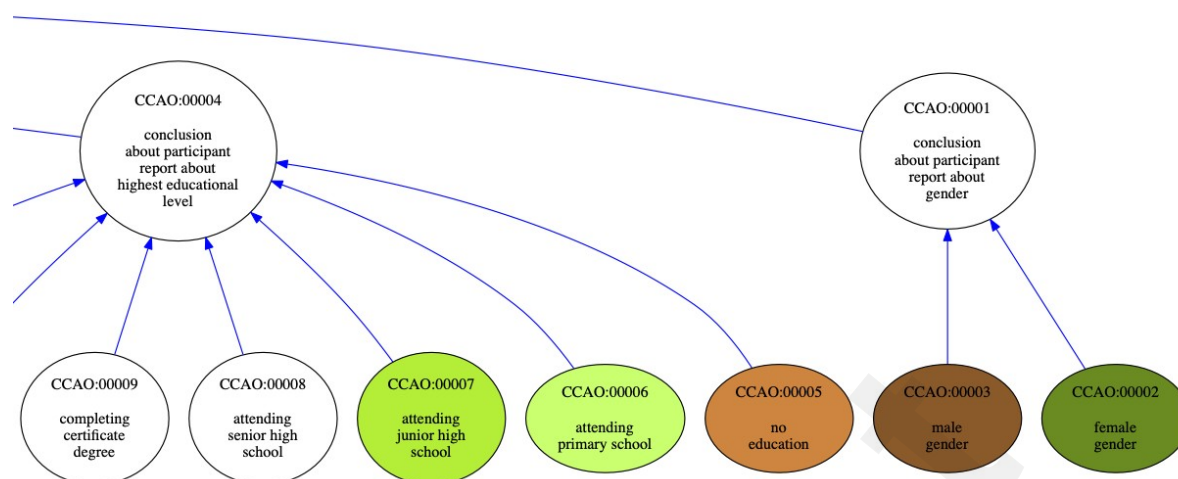


Figure 6. Zoomed-in term enrichment analysis results for subgroup of having suspected CCA among participants who underwent verbal and ultrasound screening.

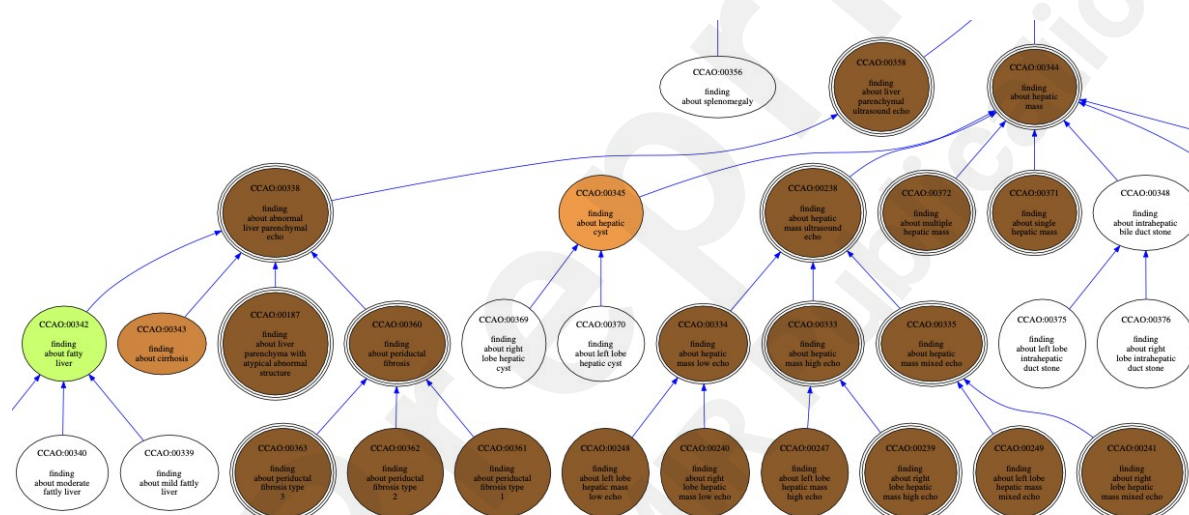


Figure 7. Zoomed-in term enrichment analysis results for subgroup of suspected CCA among participants who underwent verbal and ultrasound screening.

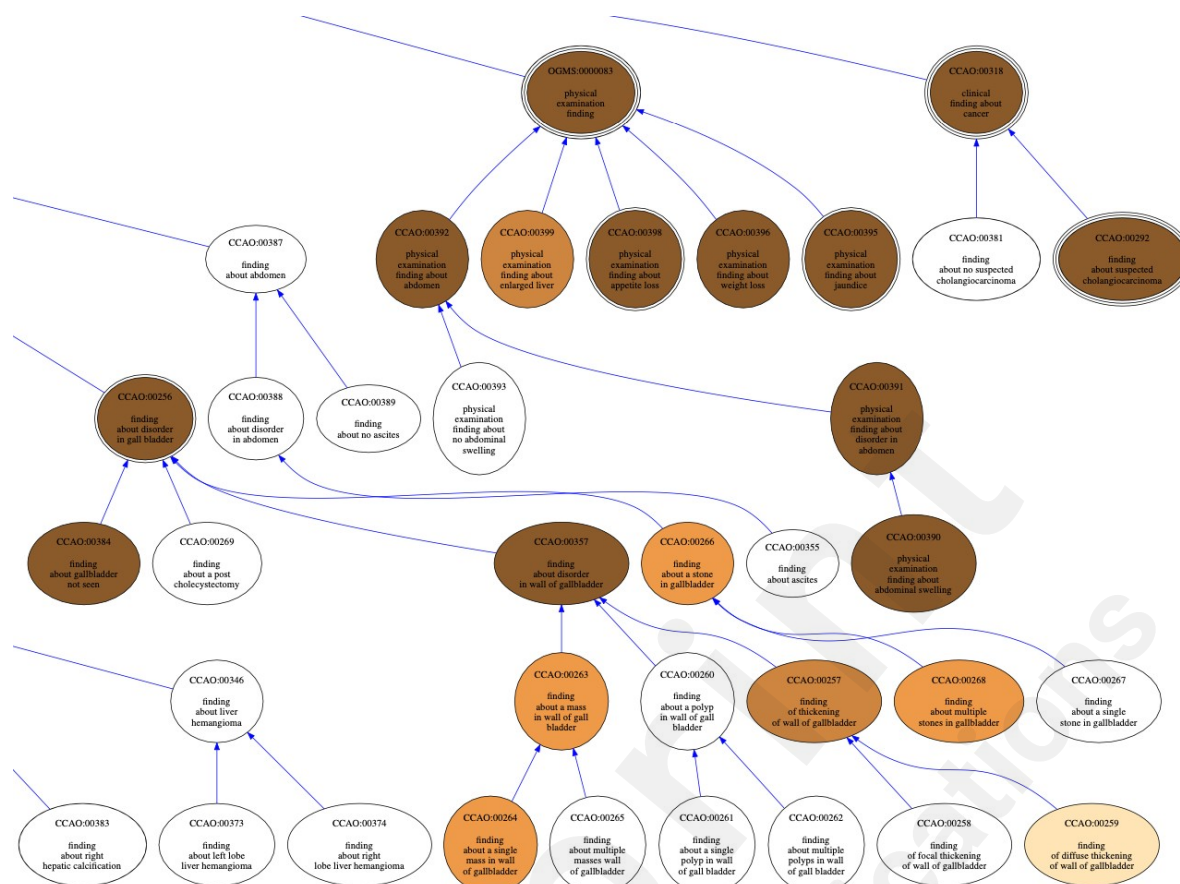


Figure 8. Zoomed-in Term enrichment analysis results for subgroup of suspected CCA (n = 380) among participants who underwent verbal and ultrasound screening.

Intrahepatic CCA patient group

For CCA, 186 participants were annotated with 'finding about intrahepatic cholangiocarcinoma' in CCAO based diagnoses in EHRs among 21,038 participants in study group having both verbal screening and ultrasound screening. In Figure 9, the term enrichment analysis result indicated annotated classes were significantly over-represented among patients with intrahepatic CCA including alcohol consumption, smoking, no education, farmer, attending primary school, completing bachelor degree, male gender, diagnosed with diabetes by clinician, diagnosed with other diseases by clinician, dilated bile duct, dilated common bile duct, dilated right lobe of bile duct, dilated left lobe of bile duct, cyst in kidney, cyst in right kidney, left kidney parenchyma with atypical abnormal function, abnormal liver parenchymal echo, liver parenchyma with atypical abnormal structure, cirrhosis, periductal fibrosis type 1, periductal fibrosis type 2, periductal fibrosis type 3, hepatic cyst, single hepatic mass, multiple hepatic mass, hepatic mass low echo, left lobe hepatic mass low echo, right lobe hepatic mass low echo, hepatic mass high echo, left lobe hepatic mass high echo, right lobe hepatic mass high echo, hepatic mass mixed echo, left lobe hepatic mass mixed echo, right lobe hepatic mass mixed echo, gallbladder not seen, mass in wall of gallbladder, single mass in wall of gallbladder, thickening of wall of gallbladder, diffuse thickening of wall of gallbladder, appetite loss, weight loss, jaundice, suspected CCA (Figure 9- Figure 16).

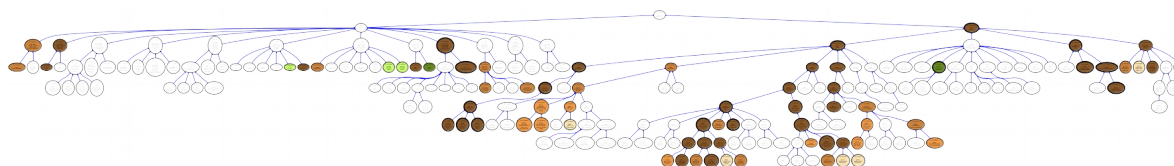


Figure 9. Term enrichment analysis results for subgroup of CCA patients among participants who underwent verbal and ultrasound screening. A red arrow indicates 'finding about suspected cholangiocarcinoma' class.

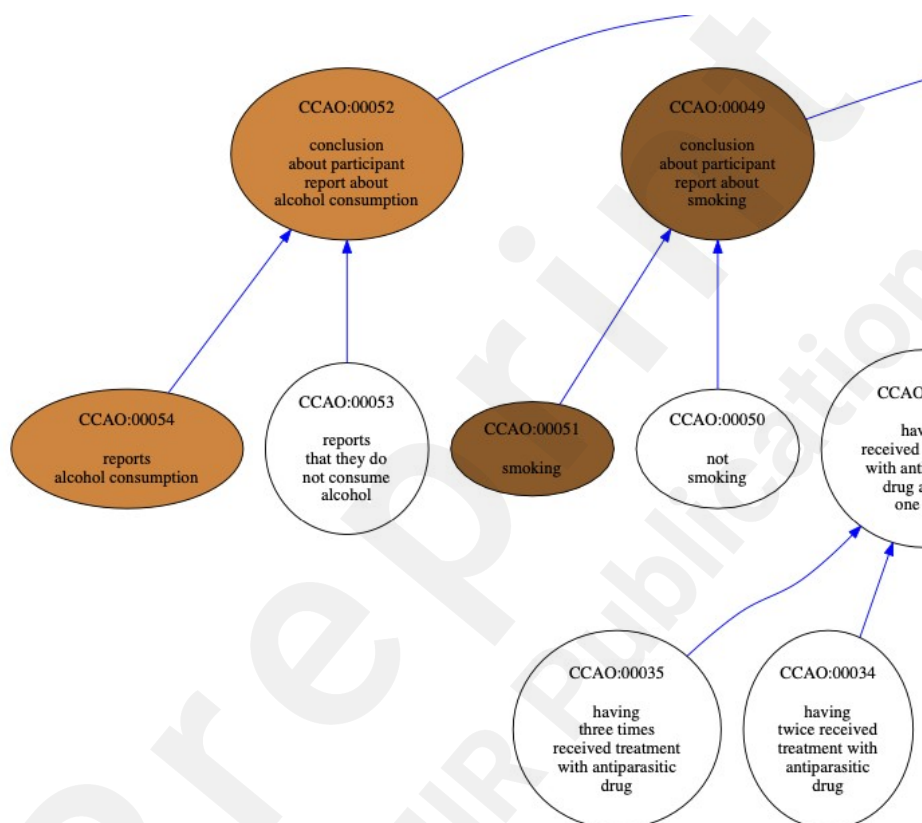


Figure 10. Zoomed-in term enrichment analysis results for the subgroup of intrahepatic CCA patients among participants who underwent verbal and ultrasound screening.

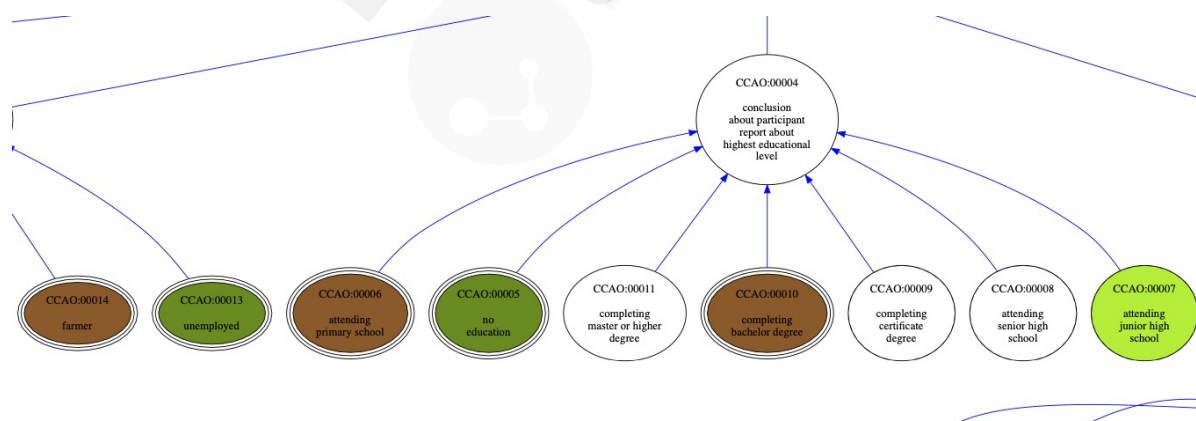


Figure 11. Zoomed-in term enrichment analysis results for the subgroup of intrahepatic CCA patients among participants who underwent verbal and ultrasound screening.

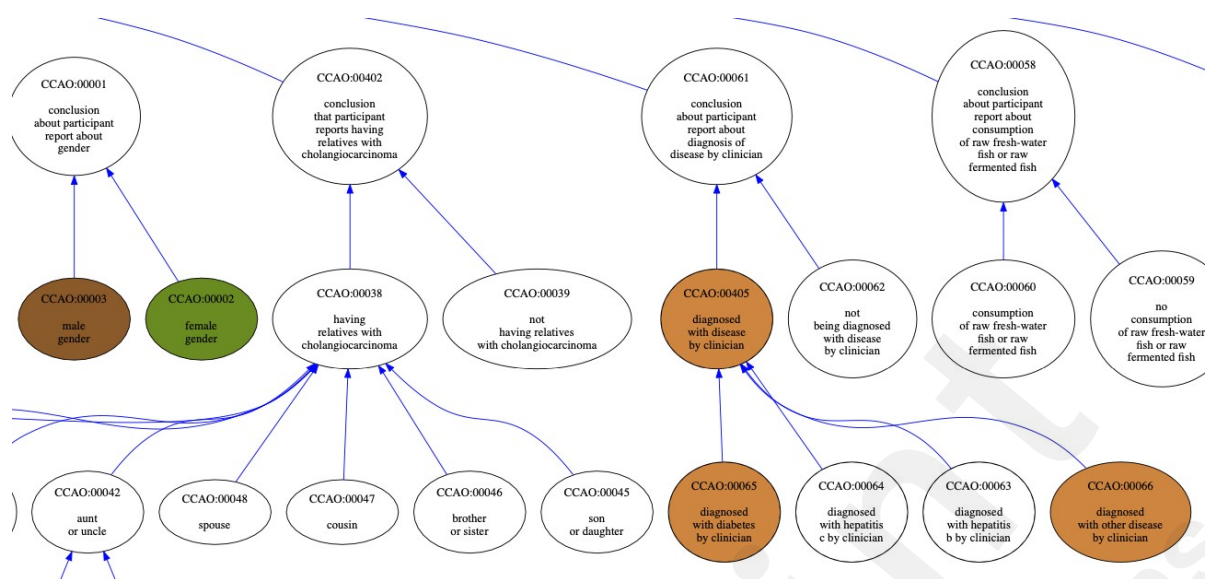


Figure 12. Zoomed-in term enrichment analysis results for the subgroup of intrahepatic CCA patients among participants who underwent verbal and ultrasound screening.

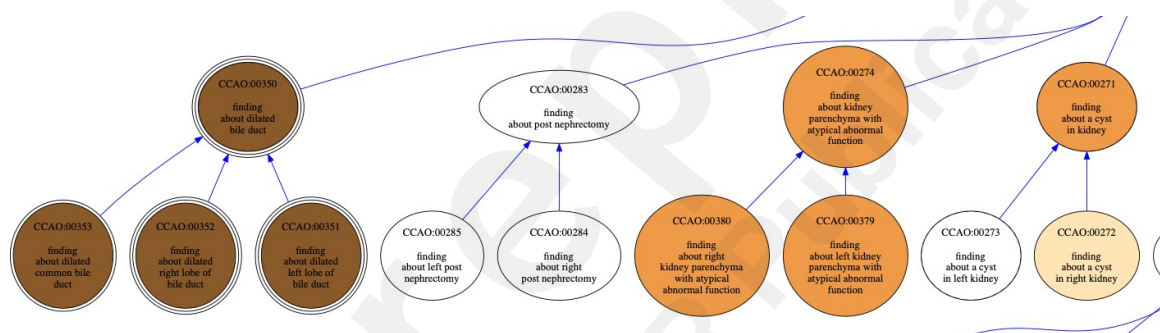


Figure 13. Zoomed-in term enrichment analysis results for the subgroup of intrahepatic CCA patients among participants who underwent verbal and ultrasound screening.

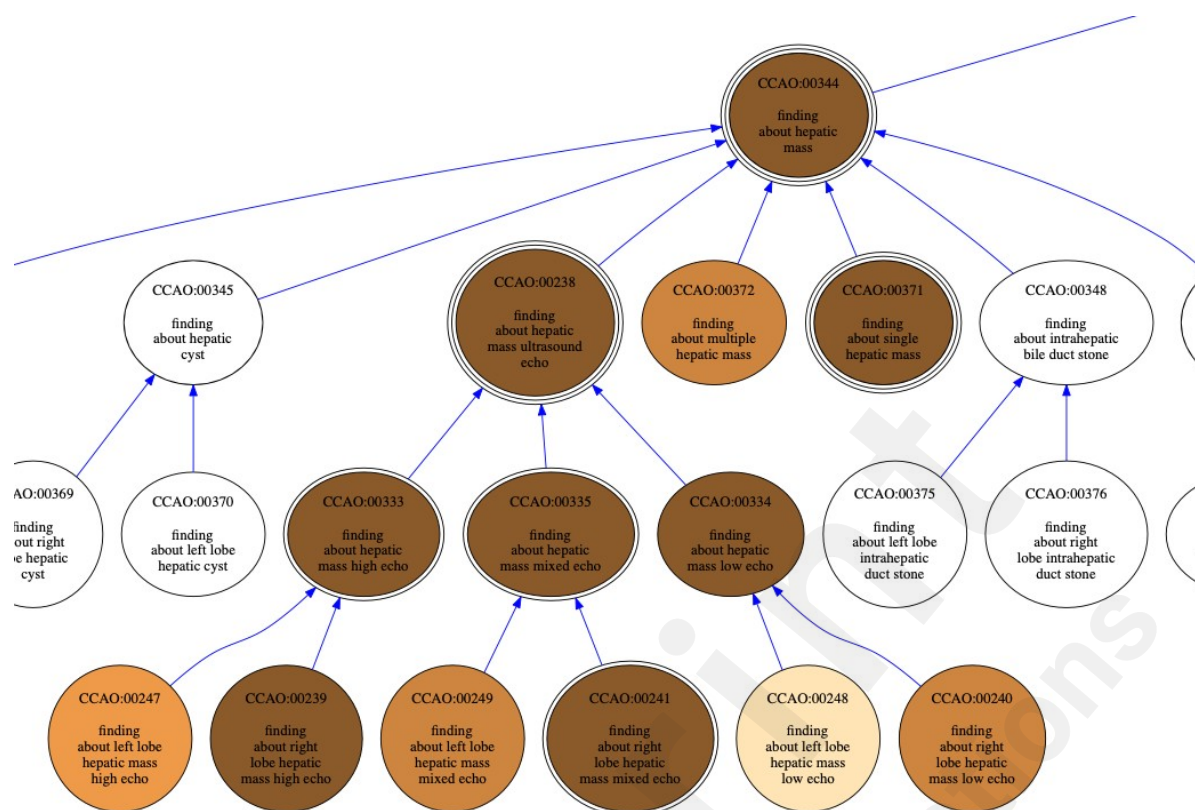


Figure 14. Zoomed-in term enrichment analysis results for the subgroup of intrahepatic CCA patients among participants who underwent verbal and ultrasound screening.

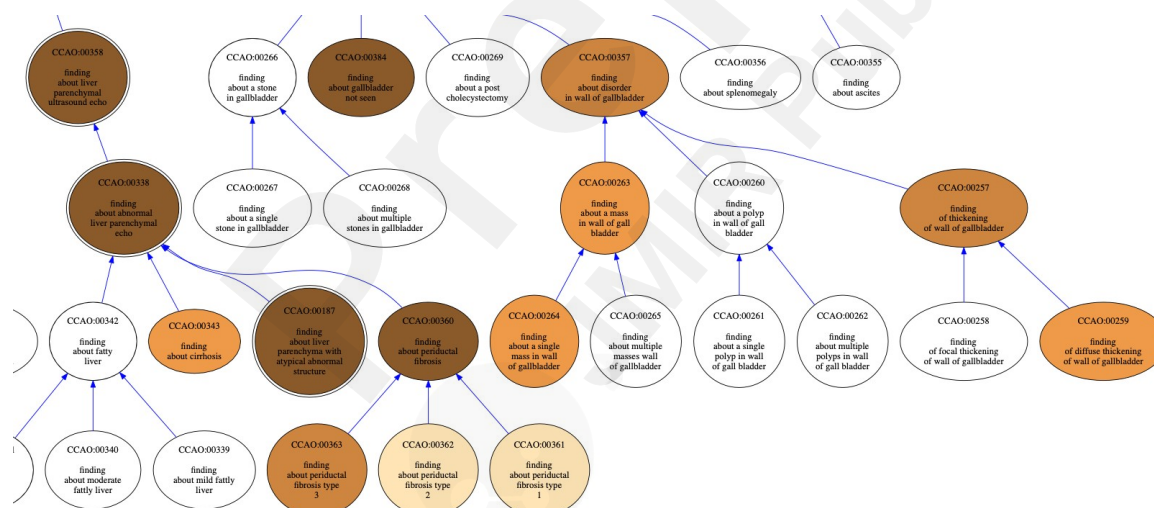


Figure 15. Zoomed-in term enrichment analysis results for the subgroup of intrahepatic CCA patients among participants who underwent verbal and ultrasound screening.

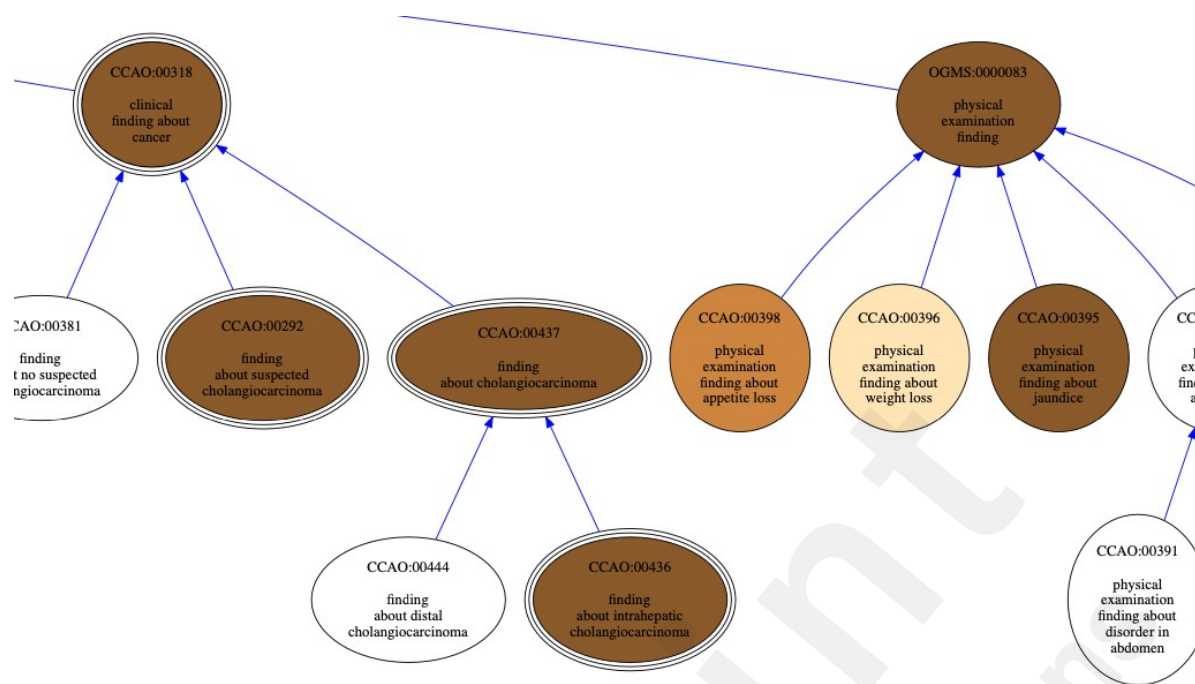


Figure 16. Zoomed-in term enrichment analysis results for the subgroup of intrahepatic CCA patients among participants who underwent verbal and ultrasound screening.

Machine learning classification for suspected CCA

We applied machine learning to the task of predicting the finding of suspected CCA using the PyCaret library, which can simultaneously run multiple machine learning algorithms. We used the over-represented variables identified in our term enrichment study as the input variables to the ML algorithms (Indicator model). The Random Forest Classifier (RF) outperformed other classification algorithms on the suspected CCA testing dataset (Table 3). In the confusion matrix, the number of true positives is 105 and of true negatives is 6,194, while the number of false positives is 4 and of false negatives is 8 (Figure 17). The RF classifier demonstrates excellent performance in distinguishing suspected CCA and non-suspected CCA with AUC of 0.9998, and the AUPRC of also confirms the model performance with average precision of 0.99 for the imbalanced data along with accuracy of 0.9979, recall of 0.921, and F1 of 0.9741.

In order to control for effect of using random number seeds in the machine learning models [24, 25] and control for the effect of main indicator variables used in the diagnoses of suspected CCA, the means and standard deviations (SD) of 10 executions of the RF classifier using different random number seeds for the indicator and non-indicator models are reported in Table 4. In non-indicator model, we removed the main indicators including dilated bile duct, periductal fibrosis, and liver mass variables. The results showed that all model performance metrics were decreased, especially, from 0.946 to 0.700 in AUPRC and from 0.850 to 0.544 in Recall when compared to the indicator model. The main indicators clearly improve the model performances for prediction of suspected CCA. Moreover, the RF classifier reveals the top ten features of importance in the indicator model including dilated bile duct, age, dilated right bile duct, dilated common bile duct, dilated left bile duct, periductal fibrosis, liver mass, abnormal liver parenchymal echo, education, and liver parenchyma with atypical abnormal structure, respectively (Figure 18).

Table 3. Comparison of Model Performances for Predictions of Participants with Suspected CCA on the Test set

Model	Accuracy	AUC	AUPRC	Recall	Precision	F1
GBC	0.9935	0.9977	0.88	0.7193	0.9011	0.8000
LR	0.9892	0.9951	0.77	0.5351	0.8026	0.6421
RF	0.9979	0.9998	0.99	0.9211	0.9633	0.9417
LIGHTGBM	0.9976	0.9995	0.98	0.9211	0.9459	0.9333
LDA	0.9818	0.9946	0.77	0.9211	0.4976	0.6462
NB	0.9604	0.9849	0.38	1.0000	0.3132	0.4

GBC = Gradient Boosting Classifier, LR = Logistic Regression, RF = Random Forest Classifier, LIGHTGBM = Light Gradient Boosting Machine, LDA = Linear Discriminant Analysis, NB = Naive Bayes, AUC = Area Under the ROC Curve, AUPRC = Area under Precision-Recall Curve

Table 4. Means and Standard Deviations for 10 Executions of the RF Classifier for the Indicator Model and the Non-Indicator Model for Suspected CCA.

Metrics	Target	Features				Mean difference	P-value
		Indicator Model		Non-Indicator Model			
		Mean	SD	Mean	SD		
AUC	Suspected CCA	0.997	0.003	0.935	0.012	0.062	4.128e-08*
AUPRC	Suspected CCA	0.946	0.013	0.700	0.034	0.246	7.014e-10*
Accuracy	Suspected CCA	0.996	0.001	0.990	0.001	0.006	1.652e-08*
F1	Suspected CCA	0.873	0.017	0.669	0.034	0.204	4.06e-09*
Precision	Suspected CCA	0.903	0.027	0.872	0.038	0.031	0.1042
Recall	Suspected CCA	0.850	0.031	0.544	0.041	0.306	1.042e-10*

The indicator model includes all over-represented classes from term enrichment results as input features. The non-indicator model excludes dilated bile duct, PDF, liver mass, and suspected CCA variables. SD = Standard Deviation, * P -value < .05

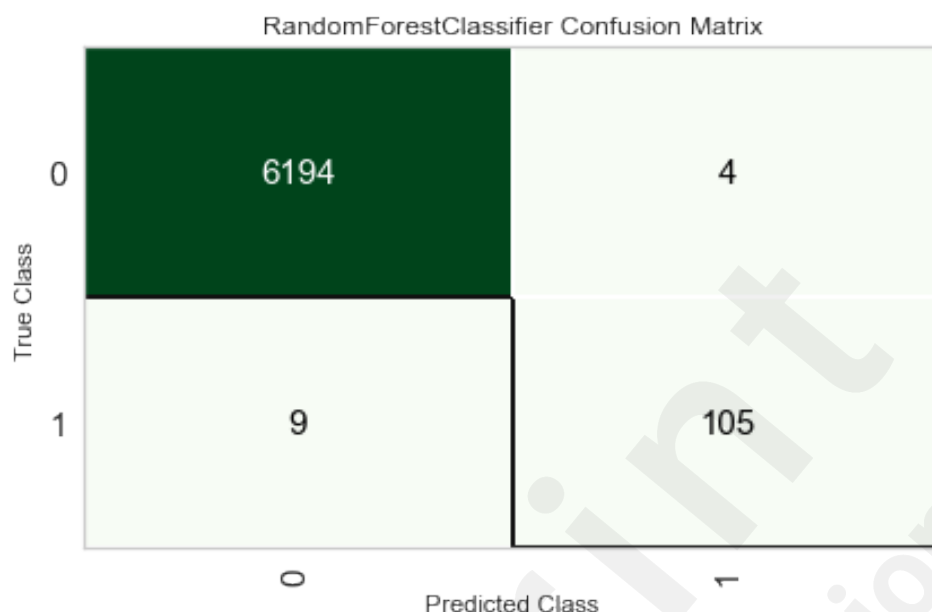


Figure 17. Random Forest Classifier Confusion Matrix for suspected CCA for the Indicator Model (0 = Non-suspected CCA, 1 = Suspected CCA).

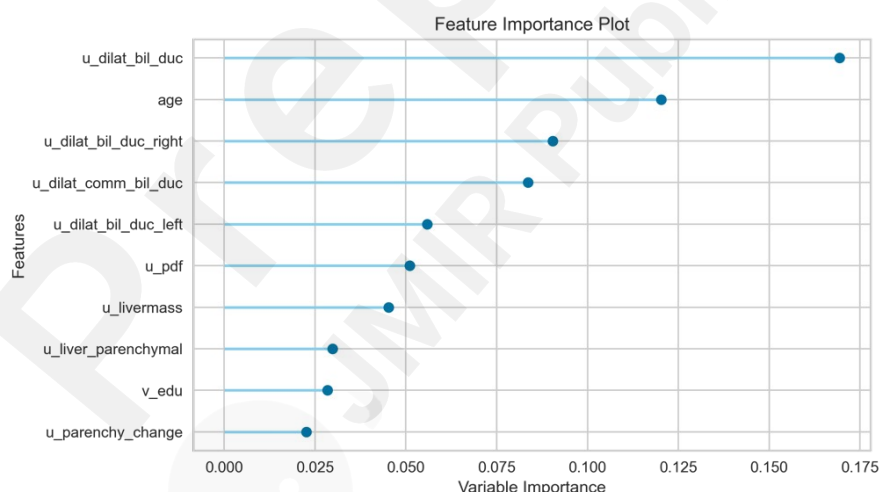


Figure 18. Feature Importance Plot of the Random Forest Classifier for suspected CCA for the Indicator Model.

Machine Learning Classification for CCA

We also applied machine learning to the task of predicting the finding of patients with CCA using the PyCaret library. We used the over-represented variables identified in our term enrichment study as the input variables to the ML algorithms (Indicator model). The Random Forest Classifier (RF) also outperformed the other classification algorithms on the CCA test set (Table 5). In the

confusion matrix, the number of true positives is 52 and of true negatives is 136, while the number of false positives is 9 and of false negatives is 6 (Figure 19). The RF classifier demonstrates excellent performance in distinguishing CCA patients and non-CCA patients with AUC of 0.9828, and the AUPRC of 0.96 also confirms the model performance for the imbalanced data along with accuracy of 0.9261, recall of 0.8966, precision of 0.8525, and F1 of 0.8739 (Table 6) dilated left bile duct, periductal fibrosis, suspected CCA, liver mass, alcohol consumption, occupation, and education, respectively (Figure 20).

As before, in order to control for effect of using random number seeds in the machine learning models [24, 25] and control for the effect of main indicator variables used in the diagnoses of patients with CCA, the means and standard deviations of 10 executions of the RF classifier with 10 different random number seeds for the indicator and non-indicator models are reported in Table 6. In non-indicator model, we removed the main indicators including dilated bile duct, periductal fibrosis, liver mass, and suspected CCA variables. The results showed that all model performance metrics were slightly decreased, for instances, from 0.920 to 0.900 in AUPRC and from 0.849 to 0.778 in Recall when compared to the indicator model. The main indicators also impact the model performances for prediction of CCA.

Table 5. Comparison of Model Performances for Predictions of CCA Patients on the Test Set

Model	Accuracy	AUC	AUPRC	Recall	Precision	F1
GBC	0.8621	0.8745	0.79	0.6552	0.8261	0.7308
LR	0.8177	0.7954	0.70	0.5345	0.7561	0.6263
RF	0.9261	0.9828	0.96	0.8966	0.8525	0.8739
LIGHTGBM	0.9261	0.9735	0.94	0.8261	0.8772	0.8696
LDA	0.8177	0.8001	0.72	0.5172	0.7692	0.6186
NB	0.7783	0.7903	0.59	0.6379	0.6066	0.6218

GBC = Gradient Boosting Classifier, LR = Logistic Regression, RF = Random Forest Classifier, LIGHTGBM = Light Gradient Boosting Machine, LDA = Linear Discriminant Analysis, NB = Naive Bayes, AUC = Area Under the ROC Curve, AUPRC = Area under Precision-Recall Curve

Table 6. Means and Standard Deviations for 10 Executions of the RF Classifier for the Indicator Model and the Non-Indicator Model for CCA Patient Prediction.

Metrics	Target	Features				Mean-Difference	P-value
		Indicator Model		Non-indicator Model			
		Mean	SD	Mean	SD		
AUC	CCA	0.950	0.021	0.937	0.022	0.013	0.03158*
AUPRC	CCA	0.920	0.023	0.900	0.026	0.02	0.04435*
Accuracy	CCA	0.920	0.013	0.906	0.021	0.014	0.0119*

Metrics	Target	Features				Mean-Difference	P-value
		Indicator Model		Non-indicator Model			
		Mean	SD	Mean	SD		
F1	CCA	0.856	0.028	0.824	0.044	0.03207	0.007462*
Precision	CCA	0.880	0.025	0.879	0.026	0.00096	0.9191
Recall	CCA	0.849	0.066	0.778	0.066	0.07115	0.00602*

Indicator models included all over-represented classes from term enrichment results as input features. While non-indicator models excluded dilated bile duct, PDF, liver mass, and suspected CCA variables. SD = Standard Deviation, * P -value<.05

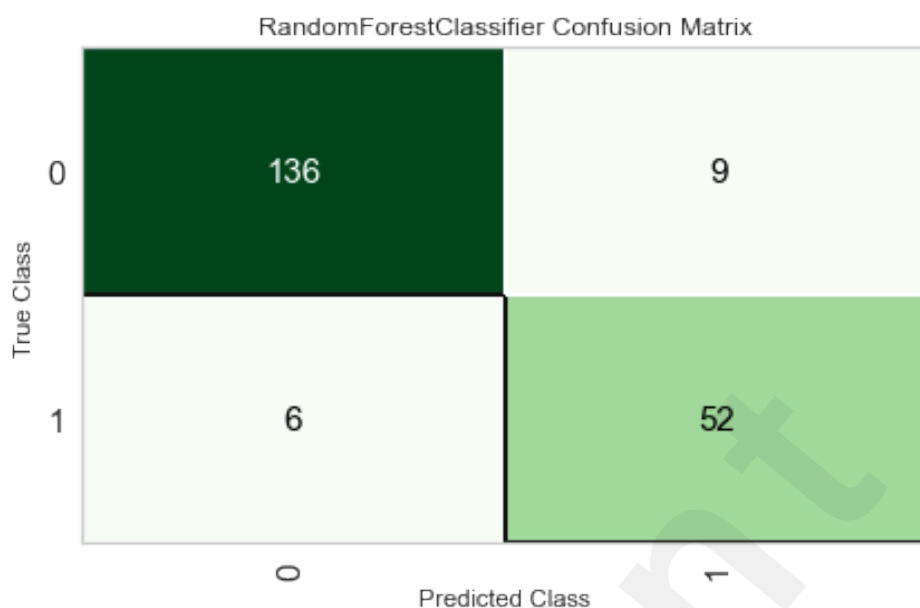


Figure 19. Random Forest Classifier Confusion Matrix for CCA on the test set (0 = non-CCA, 1 = CCA).

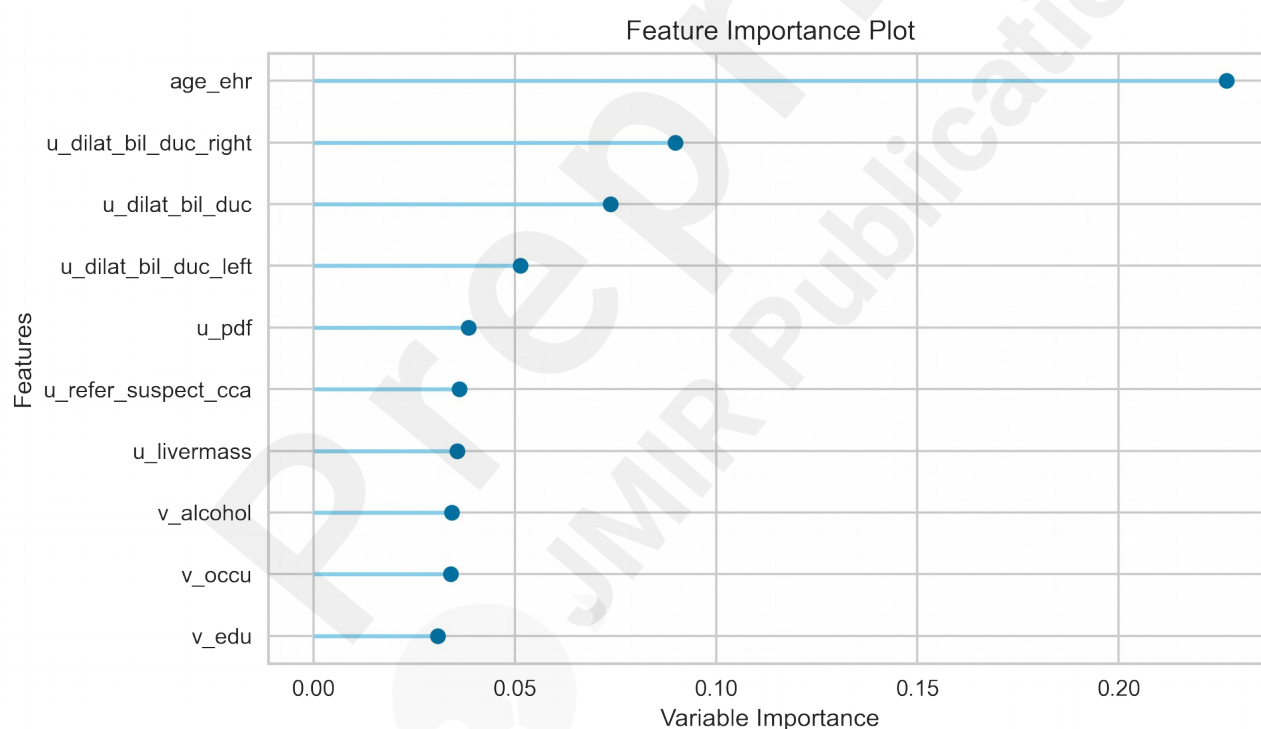


Figure 20. Feature Importance Plot of Random Forest Classifier for CCA on the test set.

Discussion

Datasets from Thailand were annotated with CCAO and analyzed using term enrichment to compare specific subpopulations to the participant group as a whole. Using term enrichment, we have identified significant risk factors in participants with suspected CCA and patients with CCA.

The results confirmed the main indicators seen in prior CASCAP results for suspected CCA

[3] including findings about having dilated bile ducts, dilated common bile duct, dilated right lobe of bile duct, dilated left lobe of bile duct, hepatic mass, hepatic mass low echo, hepatic mass high echo, hepatic mass mixed echo, periductal fibrosis, and periductal fibrosis types I, II, or III. Accordingly, Thinkhamrop et al. (2022) indicated that liver mass and dilated bile duct have significantly strong association with CCA, and histological findings of PDF among CCA patients indicates that PDF is associated with a severe bile duct disorder and CCA [26]. We found additional significant echography results that are overrepresented in patients with suspected CCA such as findings about thickening of the wall of gall bladder, and jaundice.

Moreover, our analysis revealed demographic and lifestyle risk factors that confirmed current knowledge [27, 28] based on verbal screening of participants, including reporting having alcoholism, alcohol consumption, smoking, no education, and male gender. Rayubkul et al. (2022) reported participants having diabetes were less likely to be associated with suspected CCA, while our TEA results indicate that participants with diabetes are over-represented as being associated with suspected CCA. Meta-analyses indicated that diabetes has a slightly positive relationship with CCA and may increase the development of CCA [29-31].

Furthermore, diabetes is significantly associated with fatty liver disease (FLD) which might facilitate the development of CCA because FLD is one of the pre-conditions for CCA [31]. In contrast to our TEA results, finding about fatty liver is not enriched in both participants with suspected CCA and patients with CCA. There should be more focus on patients with diabetes during and after the US screening in order to understand the relationship between diabetes and CCA better. There are very few studies focusing on suspected CCA, and more research is needed to investigate about risk factors so that we can improve the screenings.

Among CCA patients, the TEA results replicated the demographic and behavior risk factors matching the current evidence including male gender, alcohol consumption, smoking, no education, farmer, attending primary school, education, and diabetes [32-34]. However, history of *O. viverrini* infection, which is the one of the main risk factors for CCA in Thailand, is not enriched in our TEA results. Many studies that indicate that a history of *O. viverrini* infection, analyzed using data from the actual findings of fecal or specimen examinations, has a strong association with CCA [33, 35, 36]. The demographic and behavior information (verbal screening) from the dataset used in this study was obtained from self-reported questionnaires, leading to potential bias in diagnoses of *O. viverrini* infection, liver fluke treatment, and underlying diseases [27, 37]. Laboratory tests for *O. viverrini* infection are required to validate the accuracy of results from the interview data.

For ultrasonography findings, the TEA results also matched the current evidence [37-39] including having dilated bile duct, periductal fibrosis, liver parenchyma with atypical abnormal structure (liver parenchymal change), liver mass, cirrhosis, appetite loss, weight loss, and jaundice. Based on our TEA from CCAO, our approach reveals more details about anatomical locations identified using US methods, such as dilated right or left lobe of bile duct, periductal fibrosis type 1-3, hepatic mass from low, high, and mixed echo, mass in wall of gallbladder, and thickening of wall of gallbladder. This in-depth knowledge can be used to guide clinicians and healthcare professionals to monitor the high-risk populations before the development of malignant tumors.

We performed machine learning classifications using enriched terms from the TEA outputs as features for developing prediction models. In this work, the indicator model using the RF classifier outperforms in differentiating participants with suspected CCA on the test set among six classifier algorithms that were trained using the same patient cohort. A feature importance plot of the indicator model shows that age at hospital visit, dilated right bile duct, dilated bile duct, dilated right bile duct, periductal fibrosis, suspected CCA, hepatic mass, alcohol consumption, occupation, and education were the top ten most influential predictors of patients with CCA. The age variable in our model matches a previous meta-analysis reporting that older people are 7-35 times more likely to have CCA than younger people [32].

A recent study indicates that dilated bile duct has a strong relationship with CCA, followed

by hepatic mass [38]. Periductal fibrosis is a potential sign of CCA caused by *O. viverrini* infection that can be detected early by US screening [26, 40]. A referral with suspected CCA from the US screening is one of the heavily weighted risk factors in the indicator model for CCA. This finding supports that screening through ultrasonography is an efficient tool for the early detection of suspected CCA among populations residing in high-risk areas and general population CCA [38, 41]. In addition, the model performances dropped when we excluded these main indicators from the machine learning prediction models for suspected CCA and patients with CCA as illustrated in Figure 21 in order to avoid the biases of the models and reports means and Standard Deviations for ten executions of the RF Classifier for the Indicator Model and the Non-Indicator Model in order to evaluate the robustness of the models. This finding shows that ultrasonography should be used as a key screening tool for CCA in an endemic population regularly. Patients with pre-malignant lesions achieved the most benefit from US screening and can be possibly cured [42]. Moreover, our indicator model demonstrated excellent performance in predicting suspected CCA and also establishes the top five features of importance including dilated bile duct, hepatic mass and, periductal fibrosis (Figure 20), in line with the current evidence [27, 37, 38].

Studies of CCA and suspected CCA risk factors have differentiated between demographic and behavior and ultrasound findings in each study [27, 33, 38, 43-45]. Our term enrichment analysis can investigate all demographic and behavior risk factors, ultrasound findings, and electronic health records in a single analysis step, enabling researchers and health professionals to identify the overview and the combinations of the risk factors for patient cohorts simultaneously. Furthermore, the TEA approach can facilitate identifying the appropriate subsets from various features as inputs for improving machine learning model performances because the enriched terms illustrate the predictors associated with outcomes. Based on our analyses, TEA can reduce many features by excluding irrelevant features (non-enriched ontology terms), reducing unnecessary complexity in the prediction models, and the results are therefore more easily interpreted [46].

Limitations

The demographic information is based on self-reported responses to a questionnaire, and it is possible participants might not remember exactly about their medical history information such as history of liver fluke examinations and treatments, which may lead recall and measurement biases. Moreover, there may be biases due to co-morbidities such as *O. viverrini* infection, diabetes, hepatitis B, and hepatitis C [27].

Comparison with Prior Work

As noted above, some of our findings echo previous work, such as the association of diabetes with an increased risk of CCA, and the association of morphological changes in the bile duct as detected by US with later development of CCA. As noted, our work does not confirm the association of dietary habits with CCA, which may be due to the use of demographic surveys in which participants misreport their dietary habits.

In any case, our approach of using ontology-driven term enrichment coupled to machine learning analysis provides a new method for analysis of public health data not seen in earlier work.

Conclusions

Our findings can be used to focus and monitor populations at risk for suspected CCA and patients with CCA and can be validated using other statistical methods and serve as the basis for new studies. Expanding CCAO with molecular data related to CCA using ontology-driven term enrichment analysis and machine learning will help us to discover new hypotheses to decrease the

morbidity and mortality of CCA in Thailand.

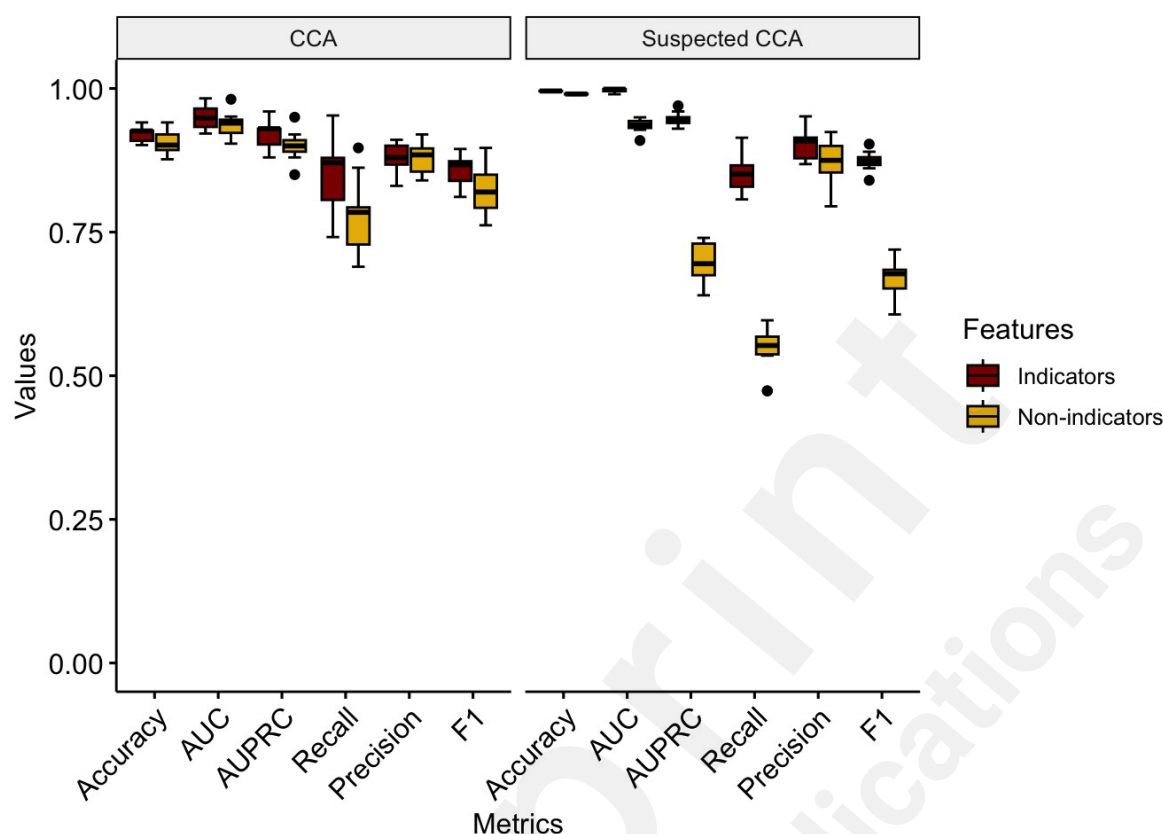


Figure 21. Mean and standard deviation of model performances for suspected CCA and patients with CCA comparing between full indicators and non-indicators models.

Acknowledgments

We would like to acknowledge Professor Werner Ceusters, of the University at Buffalo, SUNY for his valuable suggestions. We would like to thank Ms. Wiphawee Laochatrapit and Mr. Thalerngkeat Sertlert from the Kalasin Provincial Public Health Office for assisting with obtaining the data. We would like to thank Mr. Pisit Wajanasa, MSc, PhD student, UC San Diego, USA for computer programming suggestions. This work was supported by the Royal Thai Government Scholarship and the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001412 to the University at Buffalo. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of Interest

None declared

Authors' Contributions

Conceptualization: AP (lead), ADD (equal)
 Data curation: AP (lead), ADD (supporting)
 Formal analysis: AP (lead), ADD (supporting)
 Funding acquisition: AP (lead), ADD (supporting)
 Investigation: AP (lead), ADD (equal)

Methodology: AP (lead), ADD (equal)
Project administration: AP (lead), ADD (supporting)
Resources: AP (lead), ADD (supporting)
Software: AP (lead), ADD (equal)
Supervision: ADD (lead), WC (supporting)
Validation: AP (lead), ADD (equal)
Visualization: AP (lead), ADD (equal)
Writing – original draft: AP (lead), ADD (supporting)
Writing – review & editing: ADD (lead), AP (supporting)

Abbreviations

CCAO: Cholangiocarcinoma Ontology
TEA: Term Enrichment Analysis
CCA: Cholangiocarcinoma
OWL: Web Language Ontology
OGMS: Ontology for General Medical Sciences

References

1. Hughes T, O'Connor T, Techasen A, Namwat N, Loilome W, Andrews RH, et al. Opisthorchiasis and cholangiocarcinoma in Southeast Asia: an unresolved problem. *Int J Gen Med*. 2017;10:227-37. PMID: 28848361. doi: 10.2147/ijgm.S133292.
2. Sripa B, Kaewkes S, Sithithaworn P, Mairiang E, Laha T, Smout M, et al. Liver fluke induces cholangiocarcinoma. *PLoS Med*. 2007 Jul;4(7):e201. PMID: 17622191. doi: 10.1371/journal.pmed.0040201.
3. Cholangiocarcinoma Foundation of Thailand. Isan Cohort. Khon Kaen University, Thailand: CASCAP: Cholangiocarcinoma and Care Program; 2016 [July 9, 2020]; Available from: <https://cloud.cascap.in.th/>.
4. Khuntikeo N, Chamadol N, Yongvanit P, Loilome W, Namwat N, Sithithaworn P, et al. Cohort profile: cholangiocarcinoma screening and care program (CASCAP). *BMC Cancer*. 2015 Jun 9;15:459. PMID: 26054405. doi: 10.1186/s12885-015-1475-7.
5. Schulz S, Rodrigues JM, Rector A, Spackman K, Campbell J, Ustün B, et al. What's in a class? Lessons learnt from the ICD - SNOMED CT harmonisation. *Stud Health Technol Inform*. 2014;205:1038-42. PMID: 25160346.
6. Bureau of Policy and Strategy. International Statistical Classification of Diseases and Related Health Problems 10th Revision Thai Modification. Nonthaburi, Thailand: Ministry of Public Health; 2016. Available from: <http://www.thcc.or.th/ebook1/2016/mobile/index.html#p=20>.
7. Pengput A, Diehl AD, editors. Ontology Representation for Cholangiocarcinoma. ICBO-2022: International Conference on Biomedical Ontology; 2022 September 25-28; Ann Arbor, Michigan, USA: CEUR Workshop Proceedings (CEUR-WS.org); 2024.
8. Arp R, Smith B, Spear AD. Building Ontologies with Basic Formal Ontology: The MIT Press; 2015. ISBN: 0262527812.
9. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. *Summit Transl Bioinform*. 2009 Mar 1;2009:116-20. PMID: 21347182.
10. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*. 2012 2012/01/31;13(1):R5. doi: 10.1186/gb-

2012-13-1-r5.

11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*. 2000 2000/05/01;25(1):25-9. doi: 10.1038/75556.
12. LePendu P, Musen MA, Shah NH. Enabling enrichment analysis with the Human Disease Ontology. *J Biomed Inform*. 2011 Dec;44 Suppl 1(Suppl 1):S31-s8. PMID: 21550421. doi: 10.1016/j.jbi.2011.04.007.
13. The GO Consortium. GO enrichment analysis. Gene Ontology; 2024 [cited 2024 February 12, 2024]; Available from: <https://geneontology.org/docs/go-enrichment-analysis/>.
14. Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*. 2015;31(4):608-9. doi: 10.1093/bioinformatics/btu684.
15. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. *Frontiers in Physiology*. 2015 2015-December-17;6. doi: 10.3389/fphys.2015.00383.
16. Ranstam J. Multiple P-values and Bonferroni correction. *Osteoarthritis Cartilage*. 2016 May;24(5):763-4. PMID: 26802548. doi: 10.1016/j.joca.2016.01.008.
17. Shah NH, Cole T, Musen MA. Chapter 9: Analyses Using Disease Ontologies. *PLoS Computational Biology*. 2012;8(12):e1002827. doi: 10.1371/journal.pcbi.1002827.
18. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*. 2009;10(S2):S1. doi: 10.1186/1471-2105-10-s2-s1.
19. Hendricks P. anonymizer: Anonymize Data Containing Personally Identifiable Information. R package version 0222015.
20. Musen MA. The protégé project: a look back and a look forward. *AI Matters*. 2015;1(4):4–12. doi: 10.1145/2757001.2757003.
21. Jensen M, Cox AP, Ray P, Teter BE, Weinstock-Guttman B, Ruttenberg A, et al., editors. *An Ontological Representation and Analysis of Patient-reported and Clinical Outcomes for Multiple Sclerosis*. ICBO; 2014: Citeseer.
22. Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. *Graphviz— Open Source Graph Drawing Tools*. Springer Berlin Heidelberg; 2002. p. 483-4.
23. Moez Ali. PyCaret: An open source, low-code machine learning library in Python. 3.1.0 ed2020.
24. Bansal J. How to Use Random Seeds Effectively. *Toward Data Science*; 2020 [cited 2024 April 10, 2024]; Available from: <https://towardsdatascience.com/how-to-use-random-seeds-effectively-54a4cd855a79>.
25. Science O-OD. Properly Setting the Random Seed in ML Experiments. Not as Simple as You Might Imagine. *Medium*; 2019 [cited 2024 April 10, 2024]; Available from: <https://odsc.medium.com/properly-setting-the-random-seed-in-ml-experiments-not-as-simple-as-you-might-imagine-219969c84752>.
26. Chamadol N, Pairojkul C, Khuntikeo N, Laopaiboon V, Loilome W, Sithithaworn P, et al. Histological confirmation of periductal fibrosis from ultrasound diagnosis in cholangiocarcinoma patients. *Journal of Hepato-Biliary-Pancreatic Sciences*. 2014;21(5):316-22. doi: 10.1002/jhbp.64.
27. Rayubkul J, Laohasiriwong W, Thinkhamrop B, Mahato RK, Khantikeo N, S P. Prevalence of Suspected Cholangiocarcinoma Based on Ultrasonography Screening and its Associated Factors in Northeastern Thailand. *Kathmandu University School Of Medical Sciences (KUSMS)*. 2022;20(79).
28. Aye Soukhathammavong P, Rajpho V, Phongluxa K, Vonghachack Y, Hattendorf J, Hongvanthong B, et al. Subtle to severe hepatobiliary morbidity in *Opisthorchis viverrini*

endemic settings in southern Laos. *Acta Trop.* 2015 Jan;141(Pt B):303-9. PMID: 25275346. doi: 10.1016/j.actatropica.2014.09.014.

29. Jing W, Jin G, Zhou X, Zhou Y, Zhang Y, Shao C, et al. Diabetes mellitus and increased risk of cholangiocarcinoma: a meta-analysis. *Eur J Cancer Prev.* 2012 Jan;21(1):24-31. PMID: 21857525. doi: 10.1097/CEJ.0b013e3283481d89.

30. Li J, Han T, Xu L, Luan X. Diabetes mellitus and the risk of cholangiocarcinoma: an updated meta-analysis. *Prz Gastroenterol.* 2015;10(2):108-17. PMID: 26557943. doi: 10.5114/pg.2015.49004.

31. Thinkhamrop K, Khuntikeo N, Phonjitt P, Chamadol N, Thinkhamrop B, Moore MA, et al. Association between Diabetes Mellitus and Fatty Liver Based on Ultrasonography Screening in the World's Highest Cholangiocarcinoma Incidence Region, Northeast Thailand. *Asian Pac J Cancer Prev.* 2015;16(9):3931-6. PMID: 25987063. doi: 10.7314/apjcp.2015.16.9.3931.

32. Kamsa-ard S, Kamsa-ard S, Luvira V, Suwanrungruang K, Vatanasapt P, Wiangnon S. Risk Factors for Cholangiocarcinoma in Thailand: A Systematic Review and Meta-Analysis. *Asian Pac J Cancer Prev.* 2018 Mar 27;19(3):605-14. PMID: 29579789. doi: 10.22034/apjcp.2018.19.3.605.

33. Poomphakwaen K, Promthet S, Kamsa-Ard S, Vatanasapt P, Chaveepojnkamjorn W, Klaewkla J, et al. Risk factors for cholangiocarcinoma in Khon Kaen, Thailand: a nested case-control study. *Asian Pac J Cancer Prev.* 2009 Apr-Jun;10(2):251-8. PMID: 19537893.

34. Khan SA, Tavolari S, Brandi G. Cholangiocarcinoma: Epidemiology and risk factors. *Liver International.* 2019;39(S1):19-31. doi: <https://doi.org/10.1111/liv.14095>.

35. Xia J, Jiang SC, Peng HJ. Association between Liver Fluke Infection and Hepatobiliary Pathological Changes: A Systematic Review and Meta-Analysis. *PLoS One.* 2015;10(7):e0132673. PMID: 26186510. doi: 10.1371/journal.pone.0132673.

36. Prasopdee S, Rojthongpond T, Chitkoolsamphan Y, Pholhelm M, Yusuk S, Pattaraarchachai J, et al. Update on the risk factors for opisthorchiasis and cholangiocarcinoma in Thailand. *Parasites Hosts Dis.* 2023 Nov;61(4):463-70. PMID: 38043542. doi: 10.3347/phd.23032.

37. Chamadol N, Khuntikeo N, Thinkhamrop B, Thinkhamrop K, Suwannatrai AT, Kelly M, et al. Association between periductal fibrosis and bile duct dilatation among a population at high risk of cholangiocarcinoma: a cross-sectional study of cholangiocarcinoma screening in Northeast Thailand. *BMJ Open.* 2019 Mar 20;9(3):e023217. PMID: 30898798. doi: 10.1136/bmjopen-2018-023217.

38. Thinkhamrop K, Khuntikeo N, Chamadol N, Suwannatrai AT, Phimha S, Kelly M. Associations between ultrasound screening findings and cholangiocarcinoma diagnosis in an at-risk population. *Sci Rep.* 2022 Aug 6;12(1):13513. PMID: 35933509. doi: 10.1038/s41598-022-17794-9.

39. Al-Bahrani R, Abuetaf Y, Zeitouni N, Sergi C. Cholangiocarcinoma: risk factors, environmental influences and oncogenesis. *Annals of clinical and laboratory science.* 2013 Spring;43(2):195-210. PMID: 23694797.

40. Intajarnsan S, Khuntikeo N, Chamadol N, Thinkhamrop B, Promthet S. Factors Associated with Periductal Fibrosis Diagnosed by Ultrasonography Screening among a High Risk Population for Cholangiocarcinoma in Northeast Thailand. *Asian Pac J Cancer Prev.* 2016;17(8):4131-6. PMID: 27644673.

41. Khuntikeo N, Titapun A, Loilome W, Yongvanit P, Thinkhamrop B, Chamadol N, et al. Current Perspectives on Opisthorchiasis Control and Cholangiocarcinoma Detection in Southeast Asia. *Front Med (Lausanne).* 2018;5:117. PMID: 29765958. doi: 10.3389/fmed.2018.00117.

42. Sungkasubun P, Siripongsakun S, Akkarachinorate K, Vidhyarkorn S, Worakitsitatorn A,

- Sricharunrat T, et al. Ultrasound screening for cholangiocarcinoma could detect premalignant lesions and early-stage diseases with survival benefits: a population-based prospective study of 4,225 subjects in an endemic area. *BMC Cancer*. 2016;16(1). doi: 10.1186/s12885-016-2390-2.
43. Suwannatrai AT, Thinkhamrop K, Clements ACA, Kelly M, Suwannatrai K, Thinkhamrop B, et al. Bayesian spatial analysis of cholangiocarcinoma in Northeast Thailand. *Sci Rep*. 2019 Oct 3;9(1):14263. PMID: 31582774. doi: 10.1038/s41598-019-50476-7.
44. Songserm N, Woradet S, Bureelard O, Charoenbut P. Evaluation of Cholangiocarcinoma Risk and its Related Factors in Wetland Geographical Communities of Ubon Ratchathani, Thailand. *Asian Pac J Cancer Prev*. 2016;17(4):1811-5. PMID: 27221857. doi: 10.7314/apjcp.2016.17.4.1811.
45. Thinkhamrop K, Khuntikeo N, Sithithaworn P, Thinkhamrop W, Wangdi K, Kelly MJ, et al. Repeated praziquantel treatment and *Opisthorchis viverrini* infection: a population-based cross-sectional study in northeast Thailand. *Infect Dis Poverty*. 2019 Mar 20;8(1):18. PMID: 30890188. doi: 10.1186/s40249-019-0529-5.
46. Gareth James DW, Trevor Hastie, Robert Tibshirani. 1-An Introduction to Statistical Learning with Applications in Python. 1 ed 2023.