ONTOLOGY REPRESENTATION FOR CHOLANGIOCARCINOMA

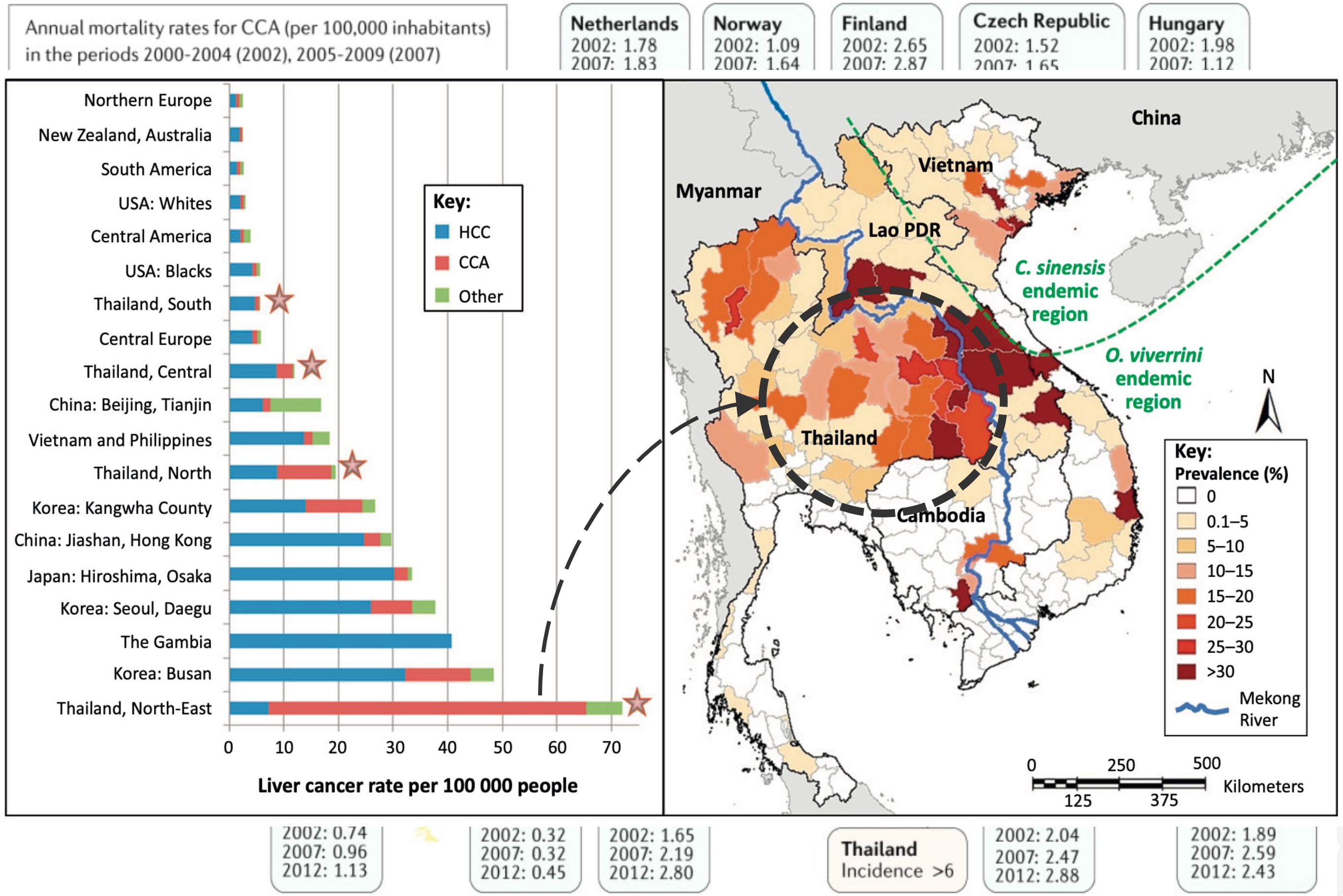
Anuwat Pengput
PhD student
Department of Biomedical Informatics
ICBO 2022, Ann Arbor, MI
September 27, 2022

UB Jacobs School of Medicine and Biomedical Sciences
University at Buffalo

Introduction

- Cholangiocarcinoma (CCA) is a major public health problem in Southeast Asia (SEA). The prevalence and incidence of CCA in Thailand and SEA are much higher than other areas in the world.
- Cultures and traditions of eating raw, fermented, pickled, and undercooked cyprinid fish (Koi Pla) are the key factor for liver fluke, *Opisthorchis viverrini (OV)*, infections.
- *OV* infections produce hepatic bile ducts and portal connective tissue inflammation.
- Chronic infections and inflammation have been indicated to be risk factors for the development of multiple stages of carcinogenesis.

(Hughes et al., 2017; Sripa et al., 2007)



Images from: Banales, 2020 DOI: 10.1038/s41575-020-0310-z
Sripa, 2012 DOI: 10.1016/j.pt.2012.07.006
Sithithaworn, 2012 DOI: https://doi.org/10.1016/j.parint.2011.08.014

Introduction (cont.)

- Several policies have been deployed over the last 40 years to prevent CCA.
- The incidence of CCA started decreasing after 2002 (Kamsa-ard et al., 2021).
- In 2015, The Cholangiocarcinoma Screening and Care Program (CASCAP), established by Khon Kaen University (KKU), Thailand, aims to eliminate OV infections and CCA. CASCAP is a prospective cohort study including the screening and patient cohorts (Khuntikeo et al., 2015).
- The cohort and electronic health records from general hospitals (Health Data Center - HDC) collect data about patient demographics and clinical phenotypes with features that have complex relationships.
- However, each database represents data elements in different ways and standards.



HDC

Health Data Center Kalasin (On Cloud)

Images from: https://cascap.kku.ac.th/, https://hdcksn.moph.go.th/

Data sources

CASCAP
CHOLANGIOCARCINOMA SCREENING AND CARE PROGRAM
Khon Kaen University

- The largest and most comprehensive databases on OVCCA
- Research-based data elements that capture details about CCA more specific than those in ICD-10-TM.
- 6 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
  - CCA-05: Post Operation Follow-up

HDC
(Health Data Center)

Heath Data Center (HDC)

- HDC is an EHR data warehouse for public hospitals.
- EHR from general and community hospitals across Thailand contain data and information on patients with or suspected of suffering from CCA including symptoms, clinical findings, treatments, and diagnoses.
- The International Statistical Classification of Disease and Related Health Problems, 10th revision, Thai Modification (ICD-10-TM) is used for morbidity and mortality coding in health services statistics, as well as for billing and payment.

Problem Statement

- Working with data from different sources and standards poses challenges of comparability.
- This work aims to address the problem of integrating and analyzing data about patients with CCA that originates from diverse sources in order to investigate risk factors for CCA.
- Data about CCA in Thailand come from public health research projects and electronic medical records in a way that make it possible to match participants in the research and patients who receive treatments in the hospitals for CCA to investigate disease outcomes.
- Combining the data from different sources requires careful application of controlled vocabulary and ontology terms to facilitate data integration and ensure understanding of its meaning.

Cholangiocarcinoma Ontology (CCAO)

- CCAO is an application ontology intended to represent data elements about cholangiocarcinoma and patient findings related to CCA.
- CCAO follows the principles of the OBO Foundry and was developed using W3C standard Web Ontology Language (OWL).
- Classes were imported from other ontologies, particularly ontologies from OBO Foundry, such as UBERON (Mungall et al., 2012).
- In the development of CCAO, we have worked to improve compatibility of imported classes with BFO.
- Many terms related to CCA do not match existing ontology classes, thus new classes were created as needed.

Overview of methods

- CCAO is being developed based on data items of the CASCAP forms which are used to collect data about demographics, ultrasound findings, diagnoses and treatments, and post-operative follow-up outcomes from targeted populations in area of Thailand where OV and CCA are endemic [4, 5].
- All variable names and data elements from the CASCAP forms were used to search on the Ontobee web browser for ontology classes [15] for matching with existing ontologies.

Development of CCAO

Translation

- CCAO based on the CCA forms 01, 02, 02.1, 03, 04, and 05
- The CCA Forms were translated from Thai to English.
- Evaluated and discussed the English versions of the CCA forms to improve the forms and ensure the correct translation.

Mapping

- Mapped data items to classes in existing ontologies using Ontobee.
- Ontology of Biological Attributes (OBA), Phenotype and Trait Ontology (PATO),
- Uberon Multi-Species Anatomy Ontology (UBERON)
- the Cell Ontology (CL),
- the Ontology of Medically Related Social Entities (OMRSE)
- Drug Ontology (DRON)

Recreating

- Created new classes based on the National Cancer Institute Thesaurus (NCIT) classes
- Defined new classes
- Assigned to the CCAO identifier (CCAO_ID)
- Referenced to the original NCIT URI using skos:closeMatch

Creating

- > 200 data items in the CCA forms that did not map to existing ontology terms.
- Did a literature review for each new ontology term
- Created new CCAO classes
- Defined new CCAO classes
- Assigned to the CCAO_ID
- Asserted logical definitions

Importing

- Created an input file (.txt) for the ROBOT tool that included the URI classes for each ontology resource
- Used ROBOT to extract classes from external ontologies and generate import files (.owl)
- All import files along with upper-level ontologies including BFO, IAO, and OGMS are imported directly to Protégé for creating CCAO
- Each new class was added manually using Protégé

Reasoning

- Removed irrelevant classes
- Manually moved classes under the upper-level ontology classes
- Checked with the ELK reasoner to prevent conflicting axioms.
- Plan to apply the CCAO CLIF-based axioms to improve the hierarchy by determining if there are any conflicting axioms in CCAO-OWL version.

Summary of ontology classes

- CCAO includes upper-level ontology classes from **BFO, OGMS, OBI, and IAO.**

Ontologies	Number of classes
CCAO classes based on CCA forms	210
CCAO classes based on NCIT classes	108
CCAO classes based on MP class	1
Uberon	23
PATO	13
OBA	8
CL	2
OMRSE	2
DRON	1

Mapping of NCIT classes

- We used 108 NCIT classes as the basis of new CCAO classes.
- These classes are needed to represent data elements on the CCA02-CCA05 forms and are classified under the top-level ontology classes including:
  - OGMS: ‘clinical finding’ such as ‘Bismuth-Corlette perihilar cholangiocarcinoma classification’, ‘cancer TNM finding’
  - OGMS: ‘disorder’ such as ‘fibrosis’, ‘cirrhosis’, ‘ascites’
  - OGMS: ‘diagnostic process’ such as ‘biopsy’, ‘computed tomography’
  - OGMS: ‘therapeutic procedure’ such as ‘cancer therapeutic procedure’ ‘percutaneous trans-hepatic biliary drainage’, ‘bypass’
  - BFO: ‘process’ such as ‘activity’, ‘referral,’ and ‘withdraw’

Why we did not use NCIT classes

- The National Cancer Institute Thesaurus (NCIT) provides comprehensive information related to CCA.
- However, we chose not to import classes from NCIT directly, because of the difficulty in merging the NCIT classes into the BFO-OGMS hierarchy.
- By creating similar classes to NCIT, we were able to write proper Aristotelian definitions and made the classes compatible with OGMS and BFO.
- This enabled us to place new subclasses of CCA within a well-formed hierarchy.

Creation of new CCAO classes

- A number of variables and data elements in the CCA forms do not match to existing ontology classes.
- We created 210 new CCAO classes along with new definitions based on data dictionary of the CCA forms and scientific literature such as

intrahepatic cholangiocarcinoma

=def. - A cholangiocarcinoma that arises from the intrahepatic bile duct epithelium in any site of the intrahepatic biliary tree.

Logical definition - *cholangiocarcinoma and (overlaps some 'intrahepatic bile duct').*

Cholangiocarcinoma hierarchy in CCAO

Differentiated by anatomical location and tumor phenotype

- cholangiocarcinoma
  - cholangiocarcinoma with non-adenocarcinoma morphology
  - distal cholangiocarcinoma
    - distal intraductal cholangiocarcinoma
    - distal mass-forming cholangiocarcinoma
    - distal periductal infiltrating cholangiocarcinoma
  - intrahepatic bile duct Intraductal papillary neoplasm with an associated invasive carcinoma
    - bile duct papillary mucin producing invasive cholangiocarcinoma
    - bile duct papillary non-mucin producing invasive cholangiocarcinoma
  - intrahepatic cholangiocarcinoma
    - intrahepatic bile duct mass-forming cholangiocarcinoma
    - intrahepatic cholangiocarcinoma of mixed type
    - intrahepatic intraductal cholangiocarcinoma
    - intrahepatic periductal infiltrating cholangiocarcinoma
  - mucin-producing cholangiocarcinoma
  - non-mucin producing cholangiocarcoma
  - non-papillary cholangiocarcinoma
    - non-papillary mucin producing cholangiocarcinoma
    - non-papillary non-mucin producing cholangiocarcinoma
  - papillary cholangiocarcinoma
    - papillary non-invasive mucin produing cholangiocarcinoma
    - papillary non-invasive non-mucin producing cholangiocarcinoma
  - papillary non-invasive cholangiocarcinoma
    - papillary non-invasive mucin produing cholangiocarcinoma
    - papillary non-invasive non-mucin producing cholangiocarcinoma
  - perihilar cholangiocarcinoma
    - Perihilar Intraductal Cholangiocarcinoma
    - Perihilar Mass-Forming Cholangiocarcinoma
    - Perihilar Periductal Infiltrating Cholangiocarcinoma
  - tubular cholangiocarcinoma
    - tubular mucin producing cholangiocarcinoma
    - tubular non-mucin producing cholangiocarcinoma
  - tubulopapillary cholangiocarcinoma
    - tubulopapillary bile duct adenocarcinoma mucin producing
    - tubulopapillary non-mucin producing cholangiocarcinoma

Logical definition of distal periductal infiltrating cholangiocarcinoma in Protégé

Description: distal periductal infiltrating cholangiocarcinoma

Equivalent To

'distal cholangiocarcinoma'
and (hasQuality some 'periductal infiltrating tumor morphology')

SubClass Of

'distal cholangiocarcinoma'

General class axioms

SubClass Of (Anonymous Ancestor)

'part of' only 'independent continuant'

'part of' only continuant

cholangiocarcinoma
and (overlaps some 'common bile duct')

neoplasm
and ('has part' some 'malignant cell')

adenocarcinoma
and (overlaps some 'bile duct')

disorder
and ('has part' some 'neoplastic cell')

Overview of the data used in this study

- Data were requested between January 2016 to May 2022, a 6 year-period, to obtain concurrent information from three datasets as much as possible including:
  1. Verbal screening dataset (CCA-01) which has demographic information, dietary information, history of O. viverrini infection and treatment, history related to CCA in participants' families and relative, and medical conditions of participants,
  2. Ultrasound screening dataset (CCA-02) which has information about clinical suspected CCA, morphology of the liver, gall stones, kidney, other findings, and transfer of patients to the hospital to confirm and treat CCA,
  3. Symptom and diagnosis (EHR) dataset which consisted of ICD-10 TM codes for symptoms, diagnoses, and causes of death related to CCA.

Enrichment Analysis

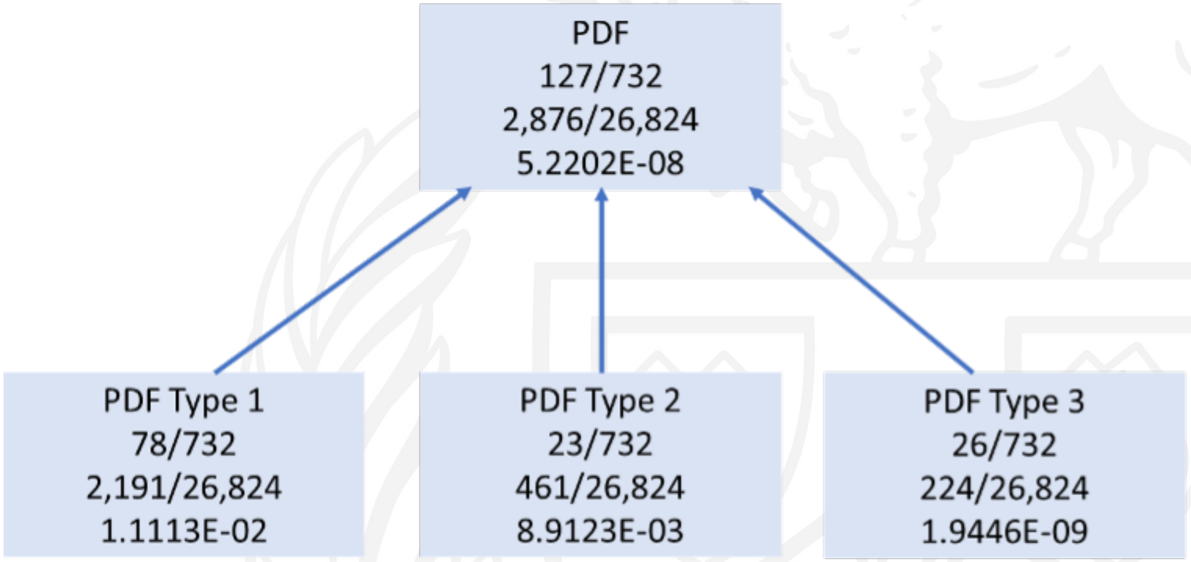
- Enrichment analyses can be done on specific populations that match certain criteria. For instance, I can select a subset of patients according to particular criteria such as age group, gender, or *O. viverrini* infection status.
- In the example here, patients with periductal fibrosis (PDF) and a later diagnosis related to CCA are compared to patients with PDF in the entire patient population.

	Success	Sample size	Percent	P-values	Bonferroni correction	Corrected P-values
PDF Type 1	78	732	10.66%	2.7781E-03	4	1.1113E-02
	2191	26,824	8.17%			
PDF Type 2	23	732	3.14%	2.2281E-03	4	8.9123E-03
	461	26,824	1.72%			
PDF Type 3	26	732	3.56%	4.8615E-10	4	1.9446E-09
	224	26,824	0.84%			
PDF Type 1+2+3	127	732	17.35%	1.3051E-08	4	5.2202E-08
	2876	26,824	10.72%			

Experiment group = participants in U/S screening with later diagnosis of CCA or symptoms related CCA
Universal group = all participants in U/S screening

Enrichment Analysis

- The results show that PDF 3 is the most enriched term (P-value = 1.9446E-09) among other PDF types. All types of PDF are statistically significant.
- This result is similar to that reported in a study by Chamadol et al., 2019, which indicated that PDF 3 significantly associated with the early stage of CCA compared with non-PDF.
- These enriched results can be used to form hypotheses that can be tested using other statistical methods.



Hierarchy of enrichment analysis among participants diagnosed with periductal fibrosis (PDF) and CCA or symptoms related to CCA compared with the whole participants in ultrasound screening

Create axioms for CCAO using CLIF

- We are developing a first-order logic (FOL) axiomatization using the Common Logic Interchange Format (CLIF) to render the CCAO fully compatible with Basic Formal Ontology 2020 (BFO2020) (Buffalo Developers Group, 2021), and to allow for reasoning that goes beyond mere classification.
- This allows for a formal integration of terms from non-BFO compatible ontologies such as SNOMED-CT by relating variables that stand for particulars also to concepts using a non-time-indexed individual-of predicate.

Create axioms for CCAO using CLIF

- For instance, we developed axioms to differentiate intrahepatic CCA from perihilar CCA and distal CCA.

```
(cl:comment "elucidation of perihilar cholangiocarcinoma"
    (forall (p t)
        (iff (instance-of p ccao-perihilar-cholangiocarcinoma t)
             (and (individual-of p sctid-cholangiocarcinoma-of-perihilar-bile-duct)
                  (instance-of p ccao-cholangiocarcinoma t)
                  (exists (h)
                    (and (instance-of h uberon-common-hepatic-duct t)
                         (overlap p h t))))))
)
```

Elucidation of perihilar cholangiocarcinoma axiom in CLIF

Conclusion

- CCAO is developed under BFO, which is now a standard for upper-level ontologies (International Standard, 2021), using best practices in ontology development.
- CCAO is publicly available on the Github repository (https://github.com/Buffalo-Ontology-Group/CCA-Ontology)
- CCAO is compatible with future expansion to represent new evidence and knowledge not be part of this initial version.
- Researchers in cholangiocarcinoma and ontology domains can reuse our ontology, and CCAO can be integrated with other ontologies because of its interoperability with its BFO compliance.
- The CCAO CLIF axiomatization will be developed to allow for reasoning and for dealing with time-indexing, negation and quantification over universals, particulars and concepts.

ACKNOWLEDGEMENTS

- Department of Biomedical Informatics
    - Werner Ceusters, MD
    - Diane G. Schwartz, MLS
    - Peter L. Elkin, MD
    - Lauren Wishnie, and Irshad Ally, MD
- Royal Thai Government Scholarship
- Ministry of Public Health, Thailand
    - Praboromarajchanok Institute,
    - Kalasin Public Health Office, Thailand
        - Ms. Wiphawee Laochaturapit, MPH
        - Mr. Thalerngkeat Sertlert
- Kavin Thinkamrop, DrPH, KKU (CASCAP), Thailand

THANK YOU