Data Anonymization and Management Pipeline for Validation of Volumetric Breast Density as an Imaging Biomarker for Predicting Breast Cancer Risk and Prognostication

by Yuhao Huang

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
B.SC. HONOURS IN DATA SIENCE
In

Irving K. Barber School of Arts and Sciences
(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA (Okanagan)
April 2020

ABSTRACT

Personalized mammography screening has emerged as the way of the future for improving breast cancer screening effectiveness. In order for individualized screening to be applicable in a clinical setting, one must first improve methods of evaluating individual risk by validating risk factors (Desreux et al. 2012). The validation of mammographic breast density as a risk factor for breast cancer requires a large, systematically anonymized and managed dataset of screening and diagnostic mammogram images. This paper describes a pipeline for anonymizing and managing mammogram data for the validation of volumetric breast density as an imaging biomarker for predicting breast cancer risk and prognostication project.

Table of Contents

List of Figures

Chapter 1 Introduction

Mammographic breast density is a promising measure to include in determining screening eligibility, as it has been shown to be a highly predictive, independent risk factor for breast cancer using 2D estimation methods. More than 40 studies have assessed the risk of breast cancer attributed to 2D breast density measured with film screen mammography and the majority reported two- to six-fold increased risk with increased breast density. However, a study conducted by Maskarinec et al recommended that ethnicity-specific models may be necessary to predict risk from breast density within different ethnic groups, as the study found Japanese women to have a lower risk than Caucasians and Native Hawaiians (Maskarinec et al. 2005). In turn, a study of women who underwent screening mammography in BC found that, while East Asian women have the highest breast density, they have a 25% less risk of developing breast cancer (Hoegg 2013). Therefore, the Early Detection group at BC Cancer Agency – Kelowna is conducting a study to incorporate ethnicity as a confounder in estimating breast cancer risk due to breast density (Rajapakshe et al. 2019).

The breast density validation will be based on screening mammogram study cases. As a result, the mammogram studies conducted at 39 study participants (34 public centers and 5 community imaging clinics) across BC are collected. Due to the large scale of data needed for the project, this study is designed for creating a pipeline to anonymize and manage the mammogram data obtained.

Objectives

1. Set up and administrate a ClearCanvas PACS server (BCCAPACS) dedicated for the project that serves as a DICOM host storing all un-anonymized, raw and processed mammogram data from study participants.
2. Implement BDen-StudyMover, a program written in C# which moves DICOM data from one workstation to another.
3. Implement BDen-StudyGrabber, a program written in C# which queries and downloads data from BCCAPACS to a workstation. Anonymization of DICOM files is done while downloading.
4. Build and maintain a MySQL database, containing records of all downloaded data.

Chapter 2 Background

2.1 Key Terms and Definitions

DICOM

Digital Imaging and Communications in Medicine, the international standard to transmit, store, retrieve, print, process, and display medical imaging information.

PACS

Picture Archiving and Communication System, a medical imaging technology which provides economical storage and convenient access to images from multiple modalities (source machine types).

eNG Network

eHealth Network Gateway, a secure, reliable, high speed network connecting most BC health care facilities within the six health authorities.

2.2 Basic DICOM Concepts

2.2.1 DICOM Data Set and Information Entities (IE)

DICOM stores information into data sets. A data set is constructed of data elements such as patient ID, patient name, study UID, etc. DICOM data elements are categorized into groups, which usually represent an information entity (IE). An IE in DICOM represents a real-world object, such as a patient or a study. There are four basic IE types in DICOM. Figure 1 explains the relationships between the four IE types in a typical screening mammogram study.

Patient IE

The Patient IE defines the characteristics of a patient who is the subject of one or more medical studies.

Study IE

The Study IE defines the characteristics of a medical study performed on a patient. A study is a collection of one or more series of medical images, presentation states, and/or SR documents that are logically related for the purpose of diagnosing a patient. Each study is associated with exactly one Patient.

Series IE

The series IE defines the attributes that are used to group composite instances (images) into distinct logical sets by certain criteria. Each series is associated with exactly one study.

Instance (Image) IE

The Image IE defines the Attributes that describe the pixel data of an image. An image is defined by its image plane, pixel data characteristics, gray scale and/or color mapping characteristics and modality specific characteristics (acquisition parameters and image creation information). An image is related to a single series within a single study.

2.2.2 Application Entity (AE)

An Application Entity (AE) is a functional component in a system that is a user and/or provider (SCU/SCP) of one or more DICOM services. Each AE uses a unique AE Title to identify itself.

2.2.3 DIMSE Services

The DICOM Message Service Element (DIMSE) provides services for message exchange between Application Entities. The DIMSE services are grouped into DIMSE-C and DIMSE-N services. The services used in this project are DIMSE-C services. There are 5 types of DIMSE-C services.

C-STORE

The C-STORE service is invoked by an AE to request the storage of SOP Instance information by a peer AE. This service is used when the calling AE moves SOP instances to the called AE.

C-FIND

The C-FIND service is invoked by an AE to match a series of attribute strings against the attributes of the set of SOP Instances managed by a peer AE. This service is used when querying from a server or a workstation.

C-GET

The C-GET service is invoked by an AE to fetch the information for one or more SOP Instances from a peer AE, based upon the attributes supplied by the invoking AE.

C-MOVE

The C-MOVE service is invoked by an AE to move the information for one or more SOP Instances from a peer AE, to a third-party AE, based upon the attributes supplied by the invoking AE. This service is used when the invoking AE requests the called AE to move SOP instances to the destination AE.

C-ECHO

The C-ECHO service is invoked by an AE to verify end-to-end communications with a peer AE. This service is used as a ‘DICOM ping’ to test communications between two AEs.

2.3 Study Overview

Breast cancer screening mammography is performed at participating public hospitals and private imaging clinics, and the goal of this study is to collect, anonymize and manage mammogram data from study participants.

All raw and processed screening mammogram data is first sent to BCCAPACS. For the 34 public imaging centers who have direct access to the eNG network, the screening mammogram studies are sent directly to BCCAPACS after they are performed (Procedure 1A in Figure 1). For the 5 private imaging clinics without access to the eNG network, a dedicated storage computer is set up at each clinic. The mammogram studies are sent to the DICOM server on the storage computer (Procedure 1B). About every 3-6 month a new, empty hard disk is sent to the clinic to swap the old one back. After the old hard disk is received at BCCA, the studies stored in it are sent to BCCAPACS with BDen-StudyMover program (Procedure 1C).

As all screening mammography study cases over a time period are stored in BCCAPACS, they will be anonymized and downloaded to a workstation computer at BCCA using BDen-StudyGrabber program (Procedure 2). Alongside the download, the MySQL database will record information of patients, studies, series, and images from downloaded mammograms.

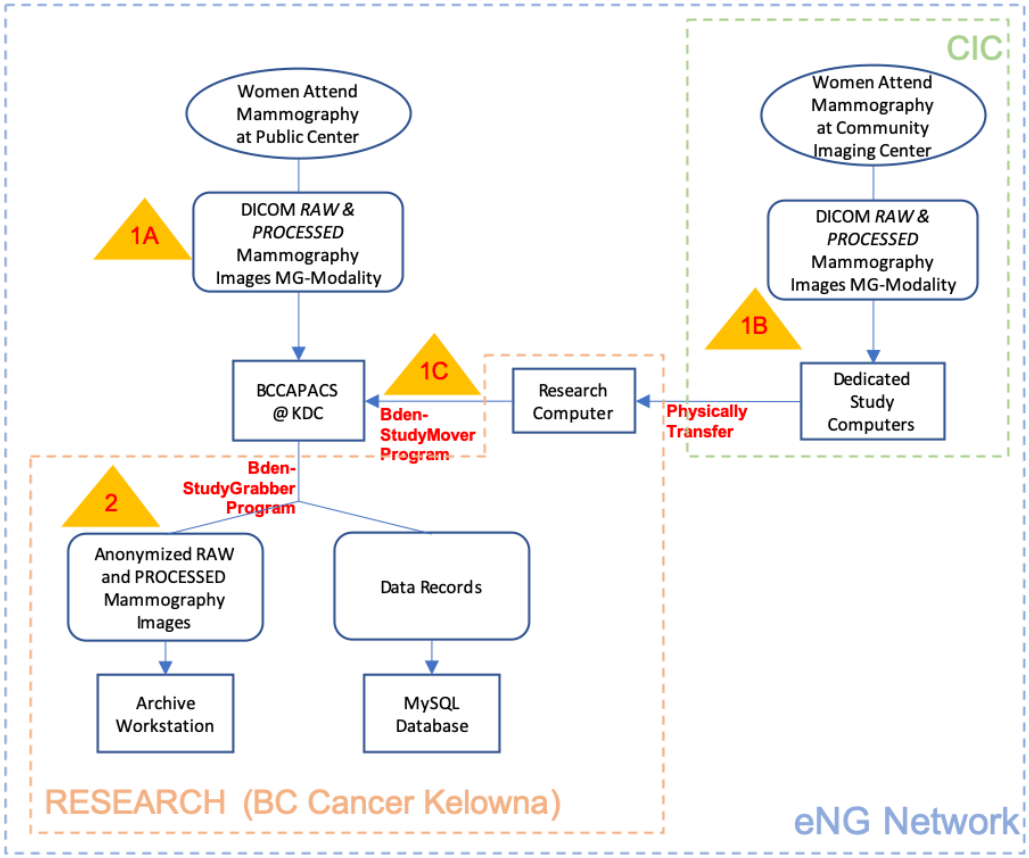

Figure 1. Data Flow Chart of the Pipeline Built in This Study

Chapter 3 BDen-StudyGrabber

BDen-StudyGrabber is a Windows form application written in C#. This program is used for querying and downloading mammogram studies from the BCCAPACS server to the workstation on which the program runs.

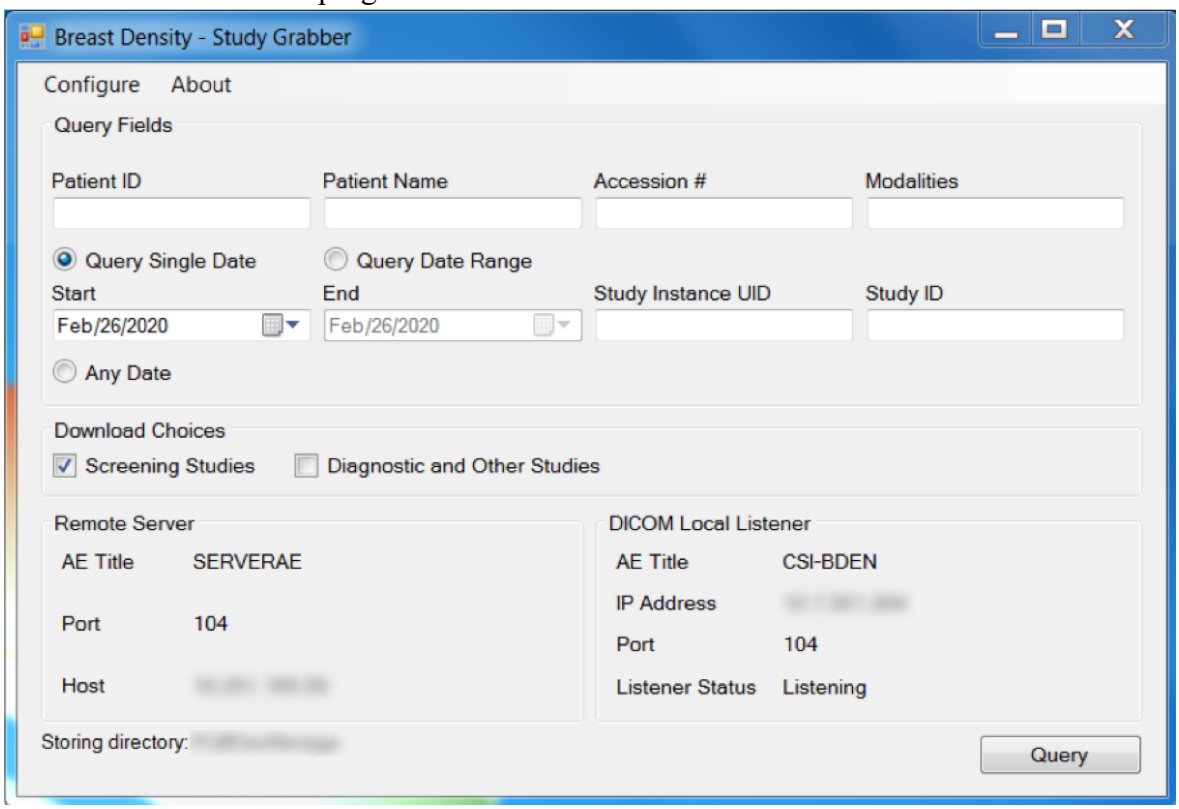

Figure 2. BDen-StudyGrabber main form

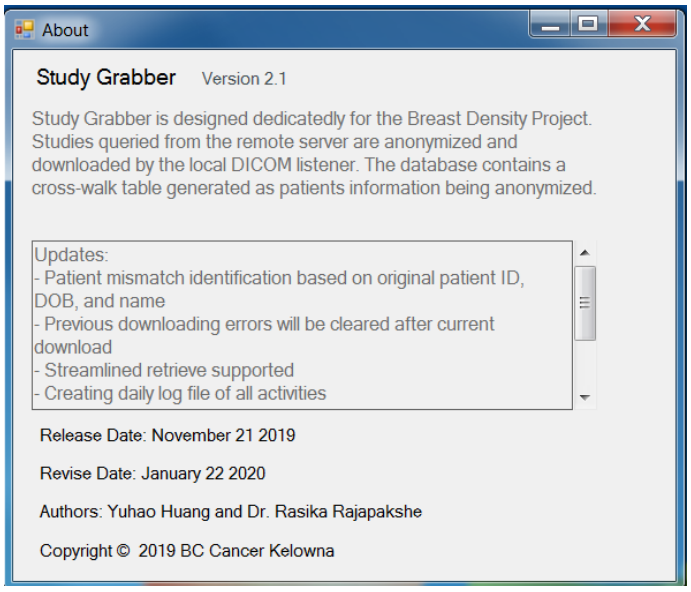

Figure 3. BDen-StudyGrabber about

3.1 Configuration

3.1.1 Configure Local DICOM Listener

BDen-StudyGrabber serves as a DICOM listener for incoming DIMSE services such as C-ECHO and C-STORE. To ensure that the program responds properly upon receiving DIMSE requests, the setting of the local DICOM listener as an AE needs to be configured, including the AE Title, port number, and IP address. Note that the IP address is always the IPv4 address of the computer on which the program is running, and therefore is not editable.

3.1.2 Configure Remote DICOM Server

The remote server is the application entity which this program will communicate with or send DIMSE services to. In this project, the remote server is the BCCAPACS.

3.1.3 Add Study Description

When downloading mammogram data, the studies are categorized as Screening, Diagnostic, QC, or Other based on their study description. When new study descriptions appear, they need to be added to the configuration in order of the program to recognize them. All configured study descriptions are saved in separate text files, one for each category, that the program reads from while downloading studies.

3.1.4 Set Storing Path

Set the path where mammogram data will be downloaded to.

3.2 Query

Before querying for mammogram files from the remote server, the user needs to specify the query criteria, that is the attribute values to match. Any attribute that remains empty will not be included in the query. The query will be done on a study level. Moreover, the user can specify the study category to download but note that the QC studies are never to be downloaded since they will not be used in the future study.

Finally, the program sends a C-FIND request to the remote server. The query result will be displayed in a new window, listing the information of the patients and studies found by the query.

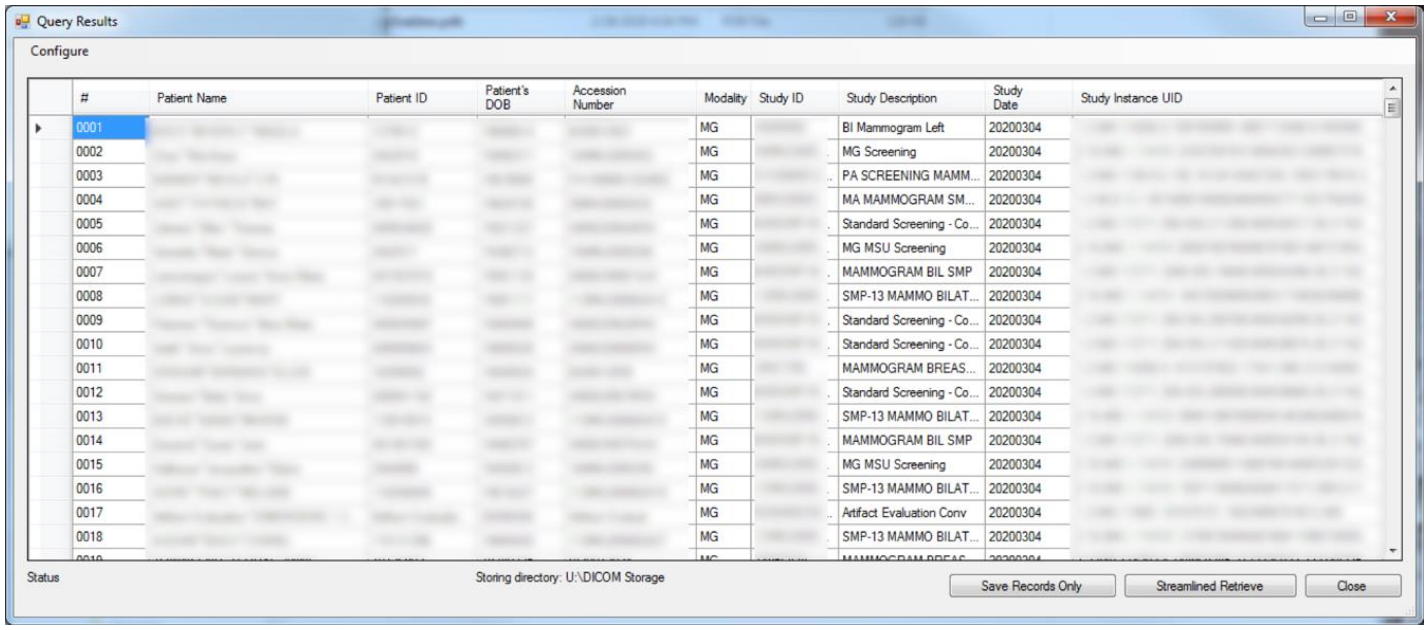

Figure 4. Query results window

3.3 Streamlined Retrieve

3.3.1 MySQL Database

The MySQL database contains information of IEs queried and downloaded. The database consists of four tables, for patients, studies, series, and images respectively. Described below are the fields in each table, and how the tables are related to each other.

Patients

1. PatientName: The patient's name, extracted from the DICOM tag Patient Name before anonymization.
2. PatientDOB: The patient's date of birth, extracted from the DICOM tag Patient's Birth Date. Note that the content of this DICOM tag is not altered during anonymization.
3. OriginalID: The patient ID assigned by the imaging center, extracted from the DICOM tag Patient's ID before anonymization.
4. BDProjectID: The patient ID assigned to the patient during anonymization (details below).

Primary keys: PatientName, PatientDOB, OriginalID.

Studies

1. StudyInstUID: The unique identifier (UID) of the study, extracted from the DICOM tag Study Instance UID. This is the primary key of the table.
2. PatientDOB: The patient's date of birth. This field is a foreign key referencing the PatientDOB field in Patients table.
3. PatientID: The project ID assigned to the patient. This field is a foreign key referencing the BDProjectID field in Patients table, indicating that the study belongs to the referenced patient.
4. Modality: The modality of the study, extracted from the DICOM tag. The value is MG for mammogram study.
5. StudyDate: The date the study is taken, extracted from the DICOM tag.
6. StudyDecription: Brief description of the study generated by the imaging center, extracted from DICOM tag.
7. AccessionNum: Accession number of the study, extracted from DICOM tag.
8. Category: Enumerated values: 'Screening', 'Diagnostic', 'QC' (Quality Control), and 'Other', indicating the purpose of the study. To determine the value of this field, the study description is compared with lists of recognized study description values of studies from configured imaging sites. If the study description is found in any list, the category is determined; if not, the study goes to the 'Other' category.
9. StudyStatus: Indicating the status of the study in the retrieving process. Enumerated values: 'Available for download', "Downloaded', and 'Error: … (error message returned from the C-MOVE request)'.

Series

1. SeriesUID: Series Instance UID extracted from the DICOM tag. This is the primary key of the table.
2. StudyUID: Study Instance UID extracted from the DICOM tag. This field is a foreign key referencing the StudyInstUID field in the Studies table, indicating that this series belongs to the referenced study.
3. SeriesDescription: Description of the series, extracted from the DICOM tag.

Images

1. SOPInstUID: The unique identifier (UID) of the image/SOP, extracted from the DICOM tag SOP Instance UID. This is the primary key of the table.
2. SeriesUID: Series Instance UID extracted from the DICOM tag. This field is a foreign key referencing the SeriesUID field in the Series table, indicating that the image/SOP belongs to the referenced series.
3. Laterality: Laterality of body part (breast) examined in the image/SOP, extracted from DICOM tag. Enumerated values: 'L' and 'R' for left and right, respectively.

4. ViewPosition: Radiographic view position in which the image/SOP is taken, extracted from DICOM tag. Enumerated values: ‘CC’ for bilateral craniocaudal and ‘MLO’ for mediolateral oblique.
5. Type: Type of mammogram image extracted from DICOM tag. Enumerated values: ‘Raw’ and ‘Processed’.
6. StoragePath: The path where the image is stored.

3.3.2 Anonymization

Prior to download, the mammogram data needs to be anonymized in order to protect patients’ personal health information (PHI). The anonymization is done by assigning a new project ID, consisting of her first screening date and a 5-digit random string, to each patient, and any DICOM tags containing PHI are replaced with the project ID. Patient IEs with the same name, date of birth, and original patient ID are considered as representing the same person, and the same project ID assigned previously is used. The information of patients and studies in the query result is written in the MySQL tables prior to download.

3.3.3 Retrieving

This is when the anonymized mammogram data is downloaded. The mammogram images are stored in the following pattern.

There is one directory for each study date named as ‘yyyymmdd’. Under the directory for the day, there is one directory for each category: Screening, Diagnostic, and Other. Note that the QC studies are not retrieved. Inside the study category directory, there is one folder for each patient who has study(ies) done on that day, named by the project ID assigned to that patient. The DICOM (.dcm) mammogram files of one study are stored in a directory for the study, named by the study date and study description, inside the folder for the patient. Each DICOM file is named by its modality, laterality, view position, and type. An example path of the DICOM file is “[root_dir]/20200101/Screening/20200101_abcde/20200101-_Screening Mammogram/MG_LCC_Raw.dcm”.

Chapter 4 BDen-StudyMover

BDen-StudyMover is a Windows form application written in C#. This program is used for querying and moving mammogram studies from the remote server to the destination. In the pipeline, this program is used for moving mammogram studies from a storage hard disk to BCCAPACS.

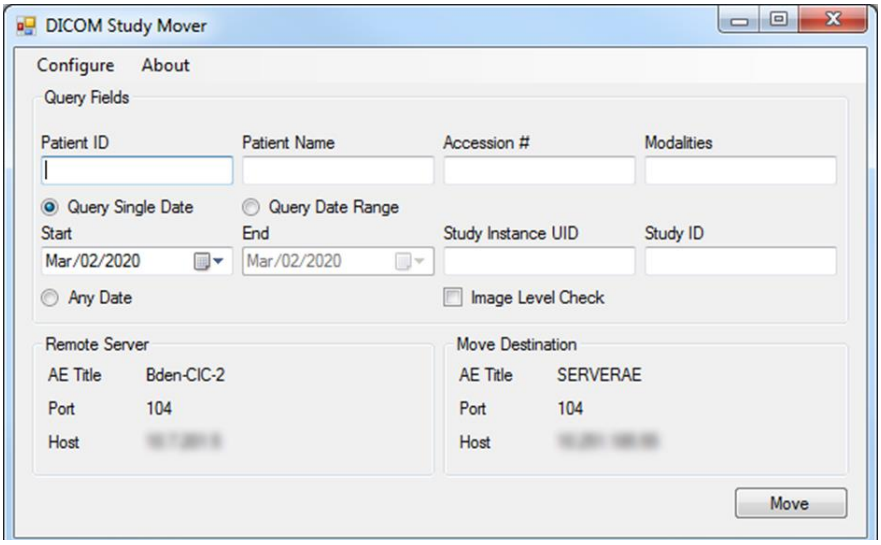

Figure 5. BDen-StudyMover main form

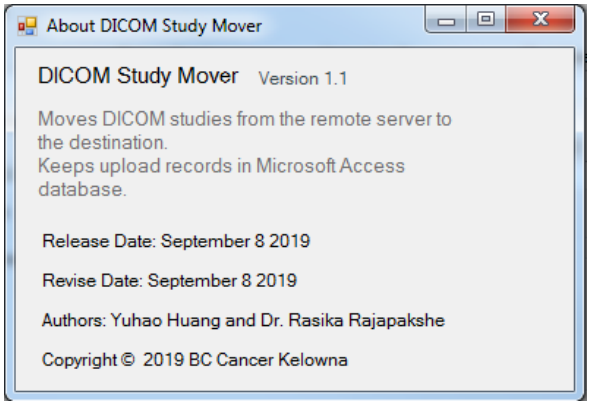

Figure 6. BDen-StudyMover about

4.1 Configuration

4.1.1 Configure Remote DICOM Server

The remote server is the application entity which this program will communicate with or send DIMSE services to. In this project, the remote server is the computer with the storage hard disks.

4.1.2 Configure Move Destination

The move destination is the application entity which the remote DICOM server will send the DICOM studies to. In this project, the move destination is BCCAPACS.

4.2 Streamlined Move

A Microsoft Access database is created by the program to keep track of the data sent from the remote server to the destination. The database consists of three tables, for the studies, series, and images to be moved respectively. Figure 7 shows the relational schema of the database. In order to insert the information of data to be moved into the database, the program first sends a study-level C-FIND request to the remote server with the query criteria specified by the user and inserts the information of each study into the studies table. The status attribute for each study is set to be 'To be moved' as this point. Secondly, a C-FIND request with the same query criteria is sent to the destination. Shall any study case in the table be found on the destination server, its status attribute will be set as 'Existed on PACS'. Thirdly, for each study case in the studies table, a series-level C-FIND request specified by the study instance UID is sent to the remote server and the information of each series is thereby inserted into the series table. Finally, an image-level C-FIND request is sent for each series UID in the series table and information of each image is inserted into the images table. At this point, the status attribute of every series and image corresponds to the status of the study case it belongs to.

For each study case in the studies table with the status of 'To be moved', a C-MOVE request with the study instance UID specified is sent to the remote server, requesting the study case to be sent to the destination server. Upon the response of each request sent, the status of the study is set as either 'Moved to PACS' or 'Error: [error message returned with the C-MOVE request]'.

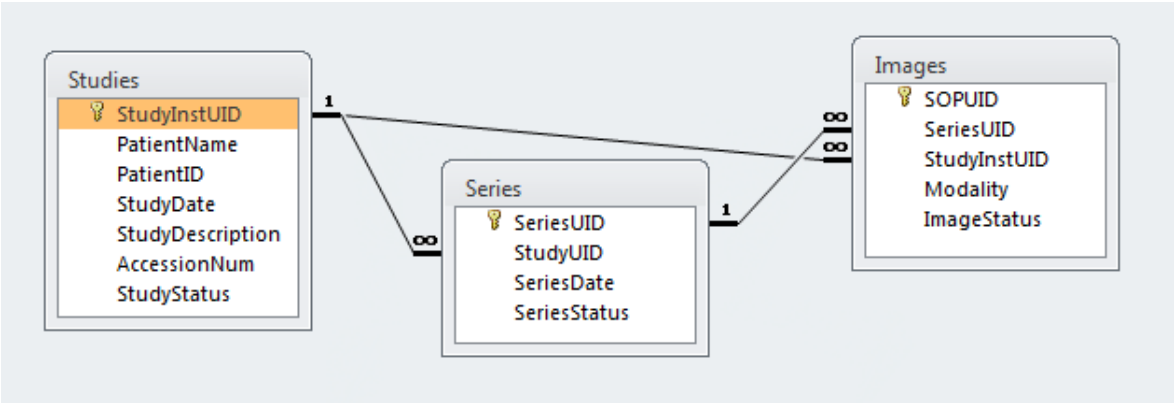

Figure 7. Relational schema diagram of the Microsoft Access database created by the BDen-StudyMover

Chapter 5 Conclusion

The pipeline described in this study is completed and is currently being used for anonymizing and downloading DICOM mammography data for the breast density validation project. By the end of March 2020, 38,700 patients and 39,390 mammography study cases of them conducted between February to June 2019 have been anonymized and downloaded，including 39,390 screening study cases.

During the development of the programs, many problems were encountered and resolved. For instance, in the early version of the BDen-StudyGrabber program, any patients with the same original patient ID are considered as the same person, and therefore are assigned with the same project ID for anonymization. However, it was brought to my attention that it was possible for different imaging centers to assign the same patient ID to different patients, as the patient IDs are not assigned by the Screening Mammography Program of BC but by the regional health authorities. Moreover, if a patient visits different imaging centers for different mammography studies, she could be assigned with different patient IDs. This problem was finally resolved by only considering the study cases with exactly same patient name, date of birth, and original patient ID to be of the same person, and any instance where one or two matches are found in the patient names, date of birth and IDs but not all the three attributes is written to a log file named 'Patient mismatch reconciliation'. During the data cleaning in the future, the researcher can review each possible mismatch case manually by reviewing the mammogram images that associate with the two patients and indicate whether more than one project ID refers to the same patient.

The pipeline can be further improved by alerting the user when study descriptions that have not appeared before are discovered and reminding the user to add them to the proper category so that study cases are not mistakenly considered as non-screening and non-diagnostic studies. Moreover, this pipeline was designed with the hope of being generalized for different DICOM modalities and research objectives.

Acknowledgement

I would like to express my sincere gratitude to Dr. Ramon Lawrence for giving me the opportunity to conduct an honours study and expertly guiding me through it, and to Dr. Rasika Rajapakshe for his generous mentoring and encouragement, without which this project would not have been possible. His motivation, sincerity and empathy have been invaluable on both an academic and a personal level, and it has been a great privilege and honour to work and study under his guidance. I would also like to thank my family and friends for the love and support.

Bibliography

Barlow, W.E., et al. “Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography.” *JNCI: Journal of the National Cancer Institute*, vol. 98, no. 17, 2006, pp. 1204–1214., doi:10.1093/jnci/djj331.

Hoegg, T. “Statistical Modelling of Breast Cancer Risk for British Columbian Women.” University of British Columbia, 2013. Web. 10 Apr. 2020. <https://open.library.ubc.ca/collections/ubctheses/24/items/1.0073888>.

Maskarinec, G., et al. “Mammographic Density and Breast Cancer Risk: The Multiethnic Cohort Study”, *American Journal of Epidemiology*, vol. 162, no. 8, 2005, pp. 743–752, https://doi.org/10.1093/aje/kwi270

Rajapakshe, R., et al. “Validation of Volumetric Breast Density as an Imaging Biomarker for Predicting Breast Cancer Risk and Prognostication”, BC Cancer Agency, 2019

DICOM Standard <https://www.dicomstandard.org/current/>