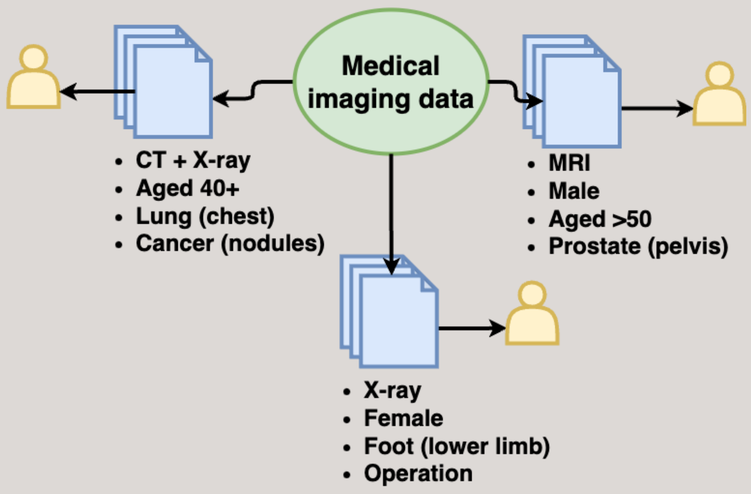
# Text-based Medical Image Classification by Body Part

Bianca Prodan, Laura Moran, EPCC (Edinburgh University); PICTURES Project (Universities of Dundee, Edinburgh); eDRIS (Public Health Scotland)

## Introduction

Research questions are usually formed around specific conditions, treatments, procedures, or demographics. Building cohorts around these questions is difficult, requiring a level of medical expertise, knowledge of the data and multiple build iterations.

To support the cohort building process, we aim to apply an automatic solution to the classification of medical images by body part. The result of this classification is labelled data which enables filtering during cohort building.



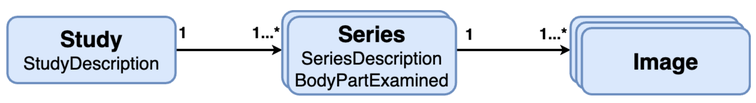

The above image shows example filters that could be applied during the cohort building process to answer specific research questions.

## 1. Text-based solution

Medical images are commonly stored as DICOM objects, containing pixel data and related metadata. DICOM metadata can describe various aspects of the image and the conditions under which it was obtained in attributes referred to as DICOM tags. DICOM tags contain patient information, scan settings, machine information, and medical notes.

Most classification solutions look at pixel data due to its high reliability in comparison with attached metadata, ignoring the metadata due to its sparseness and messiness. In comparison, text-based classification would be faster, more scalable, and more computationally efficient. Rather than seeking to replace pixel-based classification, we aim to use both the text and the image as complementary techniques to provide a confidence indicator.

To decide whether there is value in metadata text classification, a radiologist examined a collection of common DICOM tags and measured their reliability for body part classification by comparing them with their respective pixel data.




An analysis of the tag frequency and completeness paired with validation against the pixel data revealed that the "StudyDescription" tag is the primary source of body part descriptors, with "SeriesDescription" and "BodyPartExamined" as secondary sources.

## 2. Metadata preparation

The Scottish Medical Imaging (SMI) dataset contains ~25 million studies across 11 modalities. To ensure efficient labelling, the metadata was first reduced to a list of unique values by applying data cleaning and grouping, with a count of studies, series and images containing each particular value.

For each of the selected tags, a list of unique values was extracted from across the 11 modalities. These were run through the data cleaning process consisting of removal of errant spaces, special characters, and capitalisation.


This resulted in a reduction of ~25 million individual values to ~33 thousand unique values for the main tag “StudyDescription”, a reduction of ~99%.

## 3. Labelling

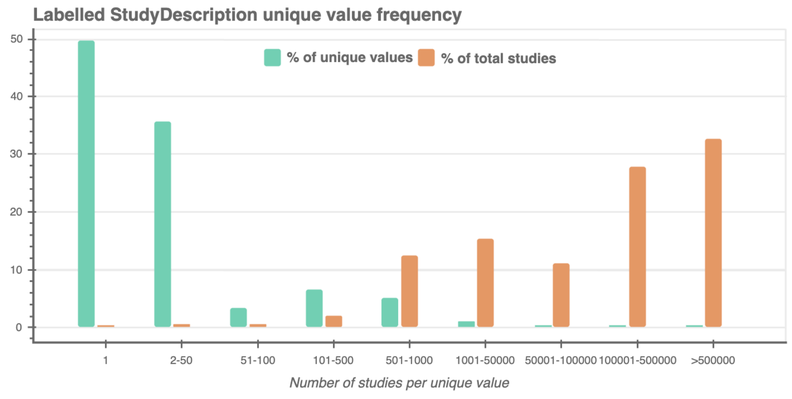
By analysing the list of unique values with a focus on the “StudyDescription” tag, we created a dictionary of common medical terms mapped to one of 9 body part categories: *head*, *neck*, *chest*, *abdomen*, *pelvis*, *upper limb*, *lower limb*, *spine*, and *whole body*. A single term can map to one or more of these categories.

The applied dictionary contained 223 terms; applying this to the unique values resulted in a coverage of ~93%. Manual verification of this coverage showed ~92% accuracy. Several issues were identified in ~4% of the unique values, including false positives, harsh abbreviations, incorrect spellings, double meanings, negations, and body part ranges, however these represented a small amount of the data.

## 4. Validation

Manual validation of the labels against the unique values as well as pixel data samples is necessary to determine the accuracy of their application.

To make this more feasible, we analysed how many of the unique values represent the majority of studies. This showed that by focusing on the top ~10% of the list we can cover ~95% of the data, with the majority of values representing single use cases.




## 5. Future Work

Labelling with a focus on the “StudyDescription“ tag acted as a proof of concept that we are currently building on by incorporating other tags and expanding the medical term dictionary.

So far, we have grouped the three selected tags: “StudyDescription“, “SeriesDescription”, and “BodyPartExamined”, labelling them separately as well as together, introducing a “confidence score” based on how many of the tags are in agreement over the applied label(s).

The confidence score provides a base for introducing other sources such as labelling from pixel-based classification solutions and would enable cohort builders to specify a minimum level of confidence when selecting data.

Once we have a stable and validated medical term dictionary, as well as labelled and unlabelled datasets, we plan on introducing Machine Learning (ML) classification methods and exploring Natural Language Processing (NLP) options for solving identified issues.

## Conclusions

While this study is in early stages, it so far shows promising results. If we can show that there is consistent value in DICOM metadata, text-based medical image classification could prove to be a valuable contribution to medical research cohort building.

Long-term viability of this solution depends on creating a scalable and automated training process for the text-based classifier. To this end, domain-specific NLP capable of dealing with medical jargon, in limited context and fragmented text sources, will need to be developed in concert with medical experts. This must be supported by a robust verification and validation process, making it a prime candidate for exploring Explainable Artificial Intelligence (XAI).

Due to the format of the data being a combination of pixel, text, potential video and audio, this work could also fit Multimodal ML models.

## Acknowledgements

University of Edinburgh

University of Dundee

Public Health Scotland (eDRIS)

PICTURES project - This work was supported by the Medical Research Council (MRC) grant No. MR/M501633/1 and the Wellcome Trust grant No. WT086113 through the Scottish Health Informatics Programme (SHIP). This project has also been supported by MRC and EPSRC (grant No. MR/S010351/1) and by the Scottish Government through the “Imaging AI” grant award.