Scottish Medical Imaging (SMI)

Susan Krueger[1], **Jackie Caldwell**[3], **Rob Wallace**[3], **Ruairidh MacLeod**[2], **Bianca Prodan**[2], **Andrew Brooks**[2], **Smarti Reel**[1], **Laura Moran**[2], **Kara Moraw**[2], **Guneet Kaur**[1], **James Sutherland**[1], **Emily Jefferson**[1] | [1]HIC (University of Dundee), [2]EPCC (Edinburgh University), [3]eDRIS (Public Health Scotland)

Introduction

Routinely collected or 'real-world' medical images linked with text-based health care records are extremely useful for healthcare research. In Scotland, all treatment is recorded under a patient's Community Health Index (CHI) number, making it easy to link images or 'scans' with clinical records spanning decades. This offers a valuable depth of data and longitudinal view of disease progression at a population level.

However, using such data is challenging because it contains personally identifiable information (PII) that must be completely removed before it can be used for research, it is also generally large and unwieldy and requires specialist tools and skills to work with. Trusted Research Environments (TREs) such as Scottish Medical Imaging (SMI) provide the capabilities and controls needed to unlock the value of such data whilst ensuring patient confidentiality.

1. Cataloguing

A large amount of metadata is attached to or can be generated for each medical image, thousands of tags depending on the image type. A Metadata Catalogue provides visibility over the SMI data, it reflects the state of data at different stages of the SMI service pipeline, providing statistics of quality, frequency, and value distributions in a single, easy-to-use catalogue.

Searching for images of a particular body part like 'chest' sounds deceptively simple, however the BodyPartExamined field is often missing or unreliable. Working with clinical experts, we automated large-scale body-part mapping from a manual mapping activity covering around 52% of CT scans, to a term-dictionary approach giving highly accurate results with between 81.03% and 99.99% coverage across 11 scan types or 'modalities'.

2. Natural Language Processing

Where labels cannot be derived from medical image metadata, these may be derived from the associated Radiology Report, or 'Structured Report' (SR).

Semantic Search System for Electronic Health Records (SemEHR) is a Natural Language Processing (NLP) tool trained on SMI data enabling search of SR free-text for specific terms and biomedical concepts from the Unified Medical Language System (UMLS) Metathesaurus.

The beauty of SemEHR is that it doesn't just do pattern matching, although it can. 'Lung Nodule' for instance can be described in different ways, these have been mapped to codes and 'concepts' in the UMLS Metathesaurus, allowing us to search by concept rather than term, and use codes from multiple ontologies for more powerful search and discovery of concepts within Radiology Reports.

Feasibility searches & Cohort identification

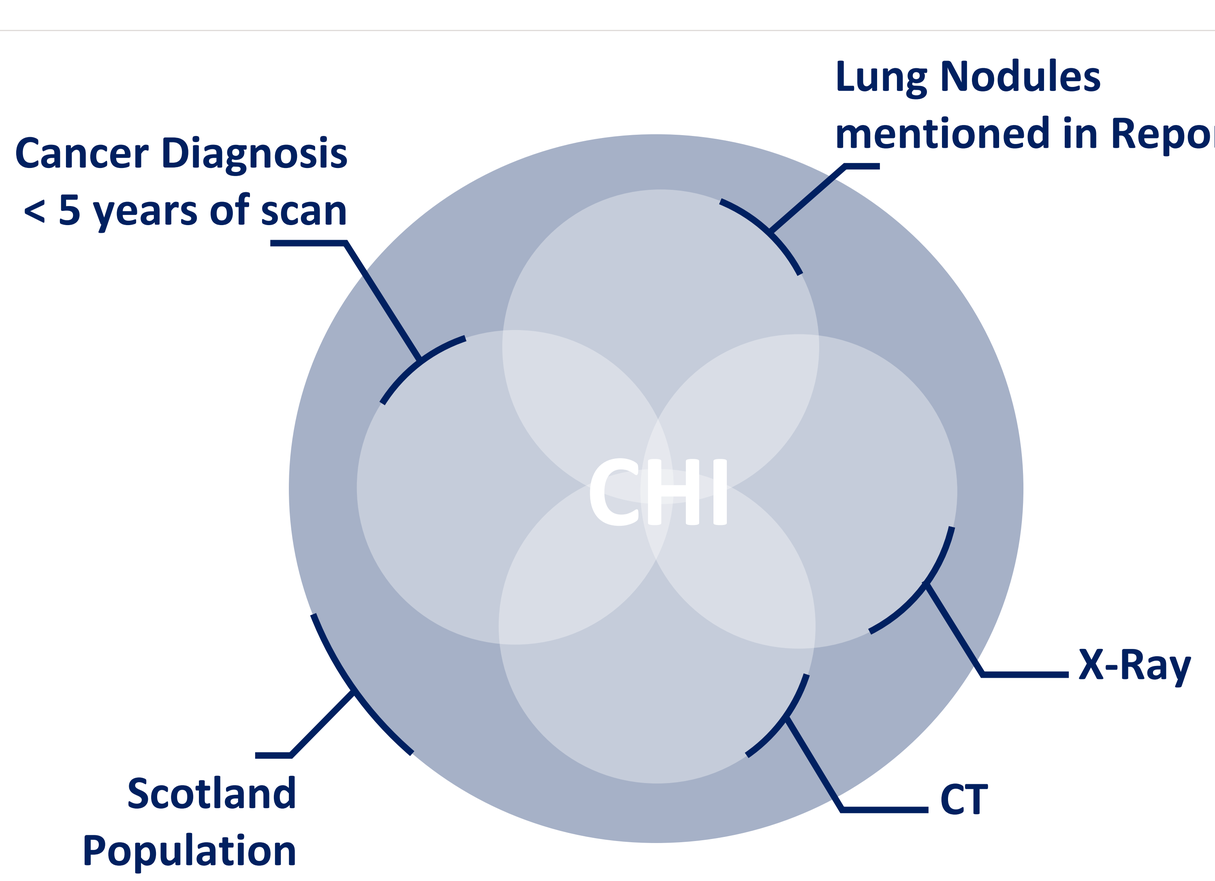
Search Criteria (example)

Inclusion:

- All Health Boards
- Conditions: Cancer diagnosis
- Age: 40 and over
- Scans: X-Ray and CT
- Body part: chest
- Indicators: Lung nodule

Exclusion:

- Scans of patient with known malignancy other than lung nodules



Matches*

Unique patients in Scotland.....4.2 m

X-Ray scans (CR & DX).....25 m

CT scans .....11 m

Mention of Lung Nodules in SR.....61 k

Patients with confirmed Lung Nodules.....35 k

Cancer diagnosis.....5 k

approximate numbers, for illustrative purposes

3. Pixel data anonymisation

Once a study cohort is confirmed, relevant linked images can be placed within a secure research environment for access by the researcher. Many scanning devices or secondary processing of images results in PII being burnt-in to the images themselves, requiring tools and methods for identifying and removing PII from pixel data at-scale. All routinely collected data must be thoroughly de-identified before it can be used for research.

We have developed and validated tools for de-identification of pixel data at scale. Modern Optical Character Recognition (OCR) tooling has been a major factor in our success as they handle grey scale very well. So well that a major challenge was over-redaction of images due to things like medical devices and teeth being picked up as text. This has been largely overcome and we continue to apply and refine the pixel de-identification tooling across modalities until all are safe enough for research.

4. Image classification

Where labels cannot be derived from medical image metadata or Radiology Reports, we seek to derive them from the images (pixel data) themselves.

Our model architecture consists of two parts: an autoencoder used to derive numerical features from the pixel data, and a classifier that is trained on those features. Both networks are trained separately. This decoupling means that the extracted features do not depend on the classification and can in fact be reused for other analyses.

This modular approach can be easily reused and expanded for further use cases, resulting in a scalable and efficient pipeline for cohort building in the context of the SMI dataset. Combined with a human-in-the-loop procedure for correcting predictions, this yields a powerful tool for deriving information from medical imaging pixel data.

Conclusions

SMI provides access to population level, research-ready medical images and associated Radiology Reports, routinely collected since 2010, linked to other health-care data. It makes them available in a secure, Trusted Research Environment or 'Safe Haven' with the tools and compute power needed for large scale analysis. Go to the Public Health Scotland website and search for Scottish Medical Imaging.

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.