Framework for large-scale automatic curation of heterogeneous cardiac MRI (ACUR MRI)

Galazis C.1; Vimalesvaran K.2; Zaman S.2; Petri C.2; Howard J.2; Linton N.2; Peters N.2; Cole G.2; Bharath AA.3; Varela M.2

1Imperial College London, Department of Computing, Faculty of Engineering, London, United Kingdom of Great Britain & Northern Ireland
2Imperial College London, Myocardial Function, National Heart and Lung Institute, Faculty of Medicine, London, United Kingdom of Great Britain & Northern Ireland
3Imperial College London, Department of Bioengineering, Faculty of Engineering, London, United Kingdom of Great Britain & Northern Ireland

Funding Acknowledgements: Type of funding sources: Public grant(s) – National budget only. Main funding source(s): UKRI CDT in AI for Healthcare http://ai4health.io and British Heart Foundation

Background
Data curation is an important process that structures and organises data, supporting research and the development of artificial intelligence models. However, manually curating a large volume of medical data is a time-consuming, repetitive and costly process that puts additional strain on clinical experts. The curation becomes more complex and demanding as more data sources are used. This leads to an introduction of disparity in the data structure and protocols.

Purpose
Here, we propose an automatic framework to curate large volumes of heterogenous cardiac MRI scans acquired across different sites and scanner vendors. Our framework requires minimal expert involvement throughout and works directly on DICOM images from the scanner or PACS. The resulting structured standardised data allow for straightforward image analysis, hypothesis testing and the training and application of artificial intelligence models.

Methods
It is broken down into three main components: anonymisation, cataloguing and outlier detection (see Figure 1). Anonymisation automatically removes any identifiable patient information from the DICOM image attributes. These data are replaced with anonymised labels, whilst maintaining relevant longitudinal information from each patient. DICOM attributes are also used to automatically group the different images according to imaging sequence (e.g. CINE, Delayed-Enhancement, T1 maps), acquisition geometry (e.g. short-axis, 2-chamber, 4-chamber) and imaging attributes (e.g. slice thickness, TE, TR), for easier querying. The sorting characteristics are flexible and can easily be defined by the user. Finally, we detect and flag, for subsequent manual inspection, any outliers within those groups, based on the similarity levels of chosen DICOM attributes. This framework additionally offers interactive image visualisation to allow users to assess its performance in real time.

Results
We tested the performance of ACUR CMRI on 26,668 CMR image series (723,531 images) from 858 patient examinations, which took place across two sites in four different scanners. With an average execution time per patient of 100 seconds, ACUR was able to sort imaging data with 1191 different sequence names into 43 categories. The framework can be freely downloaded from https://bitbucket.org/cmr-ai-working-group/acur/

Conclusions
We present ACUR, an automatic framework to curate large volumes of heterogeneous cardiac MRI data. We show how it can quickly and automatically curate data, grouping it according to desired imaging characteristics defined in DICOM attributes. The proposed framework is flexible and ideally suited as a pre-processing tool for large biomedical imaging data studies.

Abstract Figure. Overview of the ACUR MRI framework