**Background:**

The transition to electronic health records (EHRs) has displayed weaknesses in proprietary data formats and challenges in transitioning paper records to digital databases. As recently as 2014, 15% of patients had to personally bring a test result to their appointment via a CD, and as many as 5% had to repeat a procedure or test after being unable to access previous results [1]. However, after decades of transition, more than 95% of hospital’s patient medical data in the US is in the EHR pipeline [2] as well as approximately 90% of all office-based physicians [3]. While there still exist challenges in the interoperability of EHRs between different hospital systems, progress has steadily moved forward. Updated standards such as the Fast Health Interoperability Resource (FHIR) standard explicitly build in API protocols to easily transfer medical data between any two providers. Issues remain in rolling out and testing these transfer protocols for security, particularly as they deal with sensitive personal health information, one of the biggest targets for cybersecurity breaches [4].

Medical images comprise a significant portion of all medical data, with sources such as CT, MRI, ultrasound, and x-rays being more widely used in patient diagnosis and treatment around the world. These vast swaths of clinical data have opened the medical image processing field, utilizing artifical intelligence to automatically pick out the relevant fields. According to an industry-wide analysis by Nova1Advisor, the total addressable market will grow to more than $20B by 2030, representing a continuous annual growth rate of 36.9% [5]. Naturally, these machine learning algorithms need to be trained on clinical data and assessed before utilization in the clinic.

Medical images are described in a number of unique formats. Broadly, medical images can contain multiple dimensions. A single chest X-Ray has 2 dimensions of data, while a CT scan uses X-Ray “slices” to build a 3-dimensional representation. When taken over time, for example through mutliple MRI scans, these 3D volumes add an additional temporal dimension. Medical images are all made up of pixel data as an array of pixel values, with additional metadata to render the pixel data properly [6]. Common medical image data formats include DICOM, NIFTI, NRRD, MINC, and Analyze.

The conversion of medical data to its digital format, and the increase in medical imaging data, have both contributed to the explosive growth of AI in healthcare. As with any machine learning application, these algorithms rely of comprehensive and expansive swaths of data to train and test their performance. There remains pressing challenges in properly assessing these algorithms, and issues particularly in collecting and reusing medical imaging data. Widely distributing patient’s medical data demonstrates a security risk, particularly as the anonymization of imaging data is imperfect [7], [8]. There is a lack of standardization of these datasets as well, reducing the reproducibility of machine learning algorithms [9]. To overcome some of these issues, researchers have turned to General Adversarial Networks (GANs) to develop synthetic versions of clinical imaging data. These data have shown comparable performance to personal clinical data sources without the risk of patient identification due to the synthetic nature of the dataset [8].

In 2015, the FDA issued guidance that the agency will not enforce compliance with medical device data systems, medical image storage devices, and medical image communications devices. These data handling systems, which must not augment data in any way, have been deemed low risk they pose to patients and the importance they play in advancing digital health interoperability. They updated the guidance in 2019 to improve clarity but maintained the same position: hardware or software dealing with medical image data storage and transfer, and not analyzing or interpreting medical image data, are not subject to FDA regulatory requirements for medical devices [10].

### Problem Statement:

The issue of data storage with medical images derives from the complexity of the underlying data. Broadly, the data can either be raw data files, or DICOM renders. Raw files are simply stored as-is, while DICOM files go through a parser to extract relevant metadata, and then stored in a database such as PostgreSQL. The most common parser is MIRTH, an open-source repository owned by NextGen Healthcare. MIRTH enables the development of processing channels to copy the contents of certain image tags for a particular image source, for example a CT scan, to input the relevant data into the database [11]. The difficulty through this process is in relating the raw data file to the derived and rendered DICOM file. An additional important consideration is in ensuring the data file is associated with the correct patient and implementing fail safes to correct improperly linked data files [12].

Additionally, to comply with HIPAA and other global health regulations, medical data sharing including medical images must be secured before transferring to another party. This includes sending data files to a cloud server. There are two techniques to maintain data security: encrypting the data or anonymizing the data. Encrypting the images allow the raw and rendered files to be stored with all metadata attached, and only those with the private key are able to access and decrypt the data. However, the data cannot be processed remotely without compromising the private key, and the raw data can thus not be re-rendered without locally downloading the data [13]. On the other hand, anonymizing the rendered images allows them to be shared without encryption as in theory, there are no ways to tie the image to the patient’s identifying information. However, the images cannot be re-rendered as they lack the raw data to process the image. The techniques used for anonymization and de-identification have come into question, as even without the patient’s name, other information such as the date of the scan, the hospital, the machine used, or the patient’s age can be used to tie images back to their person [13]. High-resolution CT or MRI scans, or even metadata embedded in the rendered image, can also inadvertently identify the patient of a de-identified image [14].

There exist issues in linking reports to their respective images. Reports provide necessary context for medical images, such as an expert opinion on the content of the image and should be tied directly to the data it is describing. Challenges have been reported in tying multiple reports to a single image [12].

Finally, there are challenges in exchanging images between vendors. The lack of standardization in enterprise imaging pipelines is suggested to impede interoperability between these different pipelines. As such, it is often difficult or even impossible for a hospital apart of one group to exchange images digitally with a vendor in a separate system.

[1] V. Patel, “A framework for secure and decentralized sharing of medical imaging data via blockchain consensus,” *Health Informatics J.*, vol. 25, no. 4, pp. 1398–1411, Dec. 2019, doi: 10.1177/1460458218769699.

[2] M. Reisman, “EHRs: The Challenge of Making Electronic Data Usable and Interoperable,” *Pharm. Ther.*, vol. 42, no. 9, pp. 572–575, Sep. 2017.

[3] “FastStats,” Sep. 06, 2022. https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm (accessed Oct. 03, 2022).

[4] “State of the Healthcare CyberSecurity Industry.” Accessed: Oct. 03, 2022. [Online]. Available: https://www.blackbookmarketresearch.com/uploads/pdf/2020-State-of-the-Cybersecurity-Industry-F(20).pdf

[5] “AI In Medical Imaging Market Size to Hit US$ 20.18 Billion by 2030,” *BioSpace*. https://www.biospace.com/article/ai-in-medical-imaging-market-size-to-hit-us-20-18-billion-by-2030/ (accessed Oct. 03, 2022).

[6] B. Saleem, “Medical Image Formats: An Introduction,” *Ango AI*, Jul. 22, 2022. https://ango.ai/medical-image-formats-introduction/ (accessed Oct. 03, 2022).

[7] K. El Emam, S. Rodgers, and B. Malin, “Anonymising and sharing individual patient data,” *The BMJ*, vol. 350, p. h1139, Mar. 2015, doi: 10.1136/bmj.h1139.

[8] L. Fezai, T. Urruty, P. Bourdon, C. Fernandez-Maloigne, and for the Alzheimer’s Disease Neuroimaging Initiative, “Deep anonymization of medical imaging,” *Multimed. Tools Appl.*, Sep. 2022, doi: 10.1007/s11042-022-13686-2.

[9] G. Varoquaux and V. Cheplygina, “Machine learning for medical imaging: methodological failures and recommendations for the future,” *Npj Digit. Med.*, vol. 5, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41746-022-00592-y.

[10] “Medical Device Data Systems, Medical Image Storage Devices, and Medical Image Communications Devices - Guidance for Industry and Food and Drug Administration Staff,” p. 9.

[11] “Mirth Connect by NextGen Healthcare.” NextGen Healthcare, Sep. 23, 2022. Accessed: Sep. 24, 2022. [Online]. Available: https://github.com/nextgenhealthcare/connect

[12] A. J. Towbin, C. J. Roth, M. Bronkalla, and D. Cram, “Workflow Challenges of Enterprise Imaging: HIMSS-SIIM Collaborative White Paper,” *J. Digit. Imaging*, vol. 29, no. 5, pp. 574–582, Oct. 2016, doi: 10.1007/s10278-016-9897-6.

[13] S. G. Langer, “Challenges for Data Storage in Medical Imaging Research,” *J. Digit. Imaging*, vol. 24, no. 2, p. 203, Apr. 2011, doi: 10.1007/s10278-010-9311-8.

[14] D. A. Clunie, D. K. Dennison, D. Cram, K. R. Persons, M. D. Bronkalla, and H. “Rik” Primo, “Technical Challenges of Enterprise Imaging: HIMSS-SIIM Collaborative White Paper,” *J. Digit. Imaging*, vol. 29, no. 5, pp. 583–614, Oct. 2016, doi: 10.1007/s10278-016-9899-4.