**Current State:**

The transition to electronic health records (EHRs) has displayed weaknesses in proprietary data formats and challenges in transitioning paper records to digital databases. In 2014, 15% of patients had to personally bring a test result to their appointment, and as many as 5% had to repeat a procedure or test after being unable to access previous results [3]. However, after decades of transition, more than 95% of hospital’s patient medical data in the US is in the EHR pipeline [4] as well as approximately 90% of all office-based physicians [5]. 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 [6].

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% [1]. 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 [2]. Common medical image data formats include DICOM, NIFTI, NRRD, MINC, and Analyze.

“As previously mentioned, not all data of interest to the medical imaging researcher are in DICOM format; if it were, research storage needs could be met easily with an enterprise DICOM archive. However, the CT investigator may also need access to the projection sinograms, the MR researcher to the K-space Fourier signals, and the US researcher to the transducer polar or rectilinear data to name just a few possibilities. It is likely that each of these data are in a proprietary file format known only to the vendor. This reality compels a difficult choice on the investigator as will be seen.

As detailed in the “[Storage](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3056978/#Sec3)” section, encryption offers a clean solution to offsite storage *as long as there is no need to process the data offsite.* The situation is acceptable for a single site storing data offsite for later recall and research processing at the original patient care site (pursuant to Internal Review Board regulations at the site). However, a multi-center research trial that requires data processing at remote medical centers presents a much more complex challenge. In this case, the data cannot be encrypted without also preventing image processing. The answer is to anonymize the file(s) before they leave the performing site, and substitute a study identifier that can link back to the original patient demographics if required.” [10]

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 anonymiziation 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].

**Sources**

[1] : <https://www.biospace.com/article/ai-in-medical-imaging-market-size-to-hit-us-20-18-billion-by-2030/>

[2] : <https://ango.ai/medical-image-formats-introduction/>

[3] : 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](https://doi.org/10.1177/1460458218769699).

[4] : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5565131/>

[5] : <https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm>

[6] : <https://www.blackbookmarketresearch.com/uploads/pdf/2020-State-of-the-Cybersecurity-Industry-F(20).pdf>

[7] : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707567/>

[8] : <https://link.springer.com/content/pdf/10.1007/s11042-022-13686-2.pdf>

[9] : <https://www.nature.com/articles/s41746-022-00592-y>

[10] : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3056978/>