Top 5 Reasons Your Life Sciences Organization Needs Modern Medical Imaging Infrastructure

A digital transformation is underway as organizations try to reduce costs, increase operational efficiency, and accelerate innovation. Medical imaging is a rich source of patient information that is regularly leveraged to support clinical trials and clinical development. This is especially important in life sciences organizations where medical imaging is a key data type used for internal and external AI initiatives to accelerate discovery of new drug therapies and devices.

A recent IBM study reported that at least 80% of effort in AI and big data projects is linked to data preparation[1] leaving minimal time for research and analysis. Managing data at the scale needed for AI inevitably requires modern software solutions for accessing and curating large volumes of data and complex AI workflows, all while maintaining data quality, privacy and compliance. Having the right infrastructure and data management tools in place can mean all the difference in a successful transition to a data-driven R&D strategy.

Over the last several years, our team has worked extensively with pharmaceutical and medical device organizations to support their data management, processing, and collaboration needs. Through this experience, we have compiled a list of common challenges related to medical imaging data management and the top reasons why having the right infrastructure in place can greatly help organizations optimize their data processes leading to greater innovation and discovery.

Reason #1

Your data is siloed in a lot of different systems and in many different geographies. A single platform is needed to maximize the usefulness of the vast amount of data.

Life sciences organizations retrieve medical images and associated data from many sources and partners including clinical research organizations (CROs), clinical institutions, internal servers and external real world data, all hosted on unique systems. Additionally, the data from these different systems is not easily accessible, not uniformly labeled and lacks consistent quality checking.

Life sciences companies need to not only bring together data from these disparate sources, but also organize and curate it to common standards for easy querying, usage, and re-usage. The diversity and complexity of medical imaging data types adds further difficulty and expense to data management. By not aggregating disparate data into one repository, organizations can't leverage the possible discoveries to be uncovered.

We recently worked with a top 5 biopharmaceutical company that faced the challenge of migrating millions of medical images and associated data for a time-sensitive large-scale R&D project. ***By centralizing their data operations into one platform, this team was able to eliminate inefficiencies and remove human errors from their data processes by replacing their historically manual processes with a rapid and automated approach to data consolidation and curation.***

With the help of our team, data from many internal and external sources were bulk-loaded to Flywheel. Every data set that came in was automatically checked for de-identification, completeness and classified based on data type (i.e MRI structural, CT low-dose). The metadata was extracted automatically and indexed for querying within our database and an algorithm appropriate for the modality was triggered to quantify the data for quality assurance (QA). The data was routed to the appropriate project within the platform and users were provided access (with varying permissions) via the Flywheel user-interface or through command line.

The automation and streamlining of processes ensured that millions of data objects (medical images, documents, patient records, etc.) were brought into a secure, easily accessible centralized site within a fraction of the time of past data migration projects.

Reason #2

Each modality requires unique workflows for optimal performance.

Since biomedical data tends to be large, complex, and diverse, a scalable solution is needed to ensure that any medical imaging platform can handle large volumes of data and onboard many active researchers. Our life sciences customers often load diverse data sets from many modalities (MR, CT, X-ray, etc) in the range of gigabytes per study, terabytes over a cohort and petabytes in legacy systems. This amounts to massive volumes of data that require an infrastructure that can do this rapidly and efficiently.

For each modality type, automated workflows specific to that modality are mandatory, since manual processes are inefficient, time-consuming, and prone to human error. As an example, for neuroimaging, as data is ingested, algorithms are triggered to ensure data completeness followed by a conversion of the DICOM data to Nifti, a neuroimaging file format regularly used for research.

The next step is to remove the skull from the image and to run a quality assurance algorithm to quantify signal-to-noise (SNR) in the brain. This pipelined process ensures that the data is of high quality and appropriate to be used as an input for any post-processing algorithm.

In the past researchers would run their own codes on this data and would spend several hours per patient study to visually assess the data for quality and artifacts.

The ability to create unique workflows for any modality in one platform ensures that data will be managed to a common standard, ensuring sustainable and scalable processes as more data is introduced into the platform.

Reason #3

You need a cloud-scale platform for high volume computation on demand.

Large scale data analysis in medical imaging often involves the use of multiple complex algorithms to create “pipelines”, where the output of one element is the input of the next one. These pipelines are necessary for image segmentation, biomarker quantification, and synthetic data creation that is applied to hundreds and thousands of data sets. Hence, in these applications, scalable medical imaging based infrastructures are necessary to not only maintain the many algorithms and associated processing workflows but also to “burst” computation to the cloud since these pipelines regularly require many CPUs/GPUs.

Flexible deployment of pipelines reduces the IT burden to maintain these algorithms over time and promotes reproducible practices.

A processing infrastructure that can leverage local compute resources for low volume processing, combined with elastic cloud scaling for large-scale processing, is a strategy that optimizes both cost and capacity.

Our customers regularly train AI algorithms and run post-processing algorithms concurrently over many CPU and GPU nodes in the cloud for thousands of data sets, a tall order for processing data on local servers where jobs are queued one after another. Additionally, preprocessing, as described in the previous section, for light computational loads can be run on local servers to reduce cloud costs.

Reason #4

Your data infrastructure needs to support machine learning workflows and their specific requirements.

As life science companies look to machine learning to guide the future of their product development, machine learning workflows with comprehensive provenance for reproducibility and regulatory approvals are needed. Ideally, organizations want the ability to easily search and locate cohorts of data, train AI models, and run data conversion and quality assurance pipelines.

A fundamental machine learning step is to create cohorts of patient populations. Naturally, a robust, searchable framework where metadata and processes are automatically indexed and immediately available for search is necessary.

Additionally, machine learning training sets need to be generated where data access logs, curation, and processing action, either manual or automatic, is logged and tracked to establish reproducibility and audit readiness.

AI developers on our platform regularly containerize and migrate their neural networks to the Flywheel platform. Since Flywheel is a relational database, many data cohorts can be created by disease indication, age, sex, geographical location, etc. without copying the data within the platform. The cohorts and the containerized AI algorithm are accessed within the user-interface or via Flywheel's Matlab or Python software development kits (SDKs). All AI/ML training results and logs of the processing are captured on the platform and available for code troubleshooting, process improvement and for regulatory submissions.

Reason #5

You need to collaborate with internal and external partners to maximize innovation and scale efficiently.

Migrating data in the life science industry from one location to another for collaboration purposes has its share of complexities, ranging from large data transfer bottlenecks to regulatory compliance. Additionally, many companies have teams located all over the world, requiring compliance with regulatory requirements for each country or region.

Having a shared secure data repository provides researchers with a seamless resource to access data and algorithms and collaborate more effectively. Collaboration is key to accelerating healthcare innovation, but concerns around data security and privacy have often gotten in the way. With the right infrastructure in place, life sciences organizations can effectively collaborate with partners internally and externally to accelerate discovery.

Once a Flywheel instance is deployed, many researchers can access data via the Flywheel user-interface. Users will log-in via common authentication frameworks and they will access only data made available to them by a site administrator via Flywheel's permission-based controls mechanism. The users will have read-only, read-write or admin privileges on this data and can contribute their data or containerized algorithms to the same instance.

The Way Forward in Life Sciences R&D

The modern life sciences company is moving towards a "data-driven" operational model. Medical imaging plays an important role in this new paradigm as the power of diagnostic tools can greatly accelerate R&D discovery and improve clinical trial outcomes. Additional data types such as digital pathology, microscopy, and genomics are complementary additions to multimodal research, adding significant value for diagnosis of complicated diseases but also creating additional complexities to the data management process.

The integration of all data types as part of a digital transformation initiative requires an all-encompassing solution that can curate and organize large volumes of these data types, enable complex processing and AI pipelines, and provide the tools necessary to enhance collaboration across many teams and partners.

Learn more at flywheel.io

1 2016 Data Science Report https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf