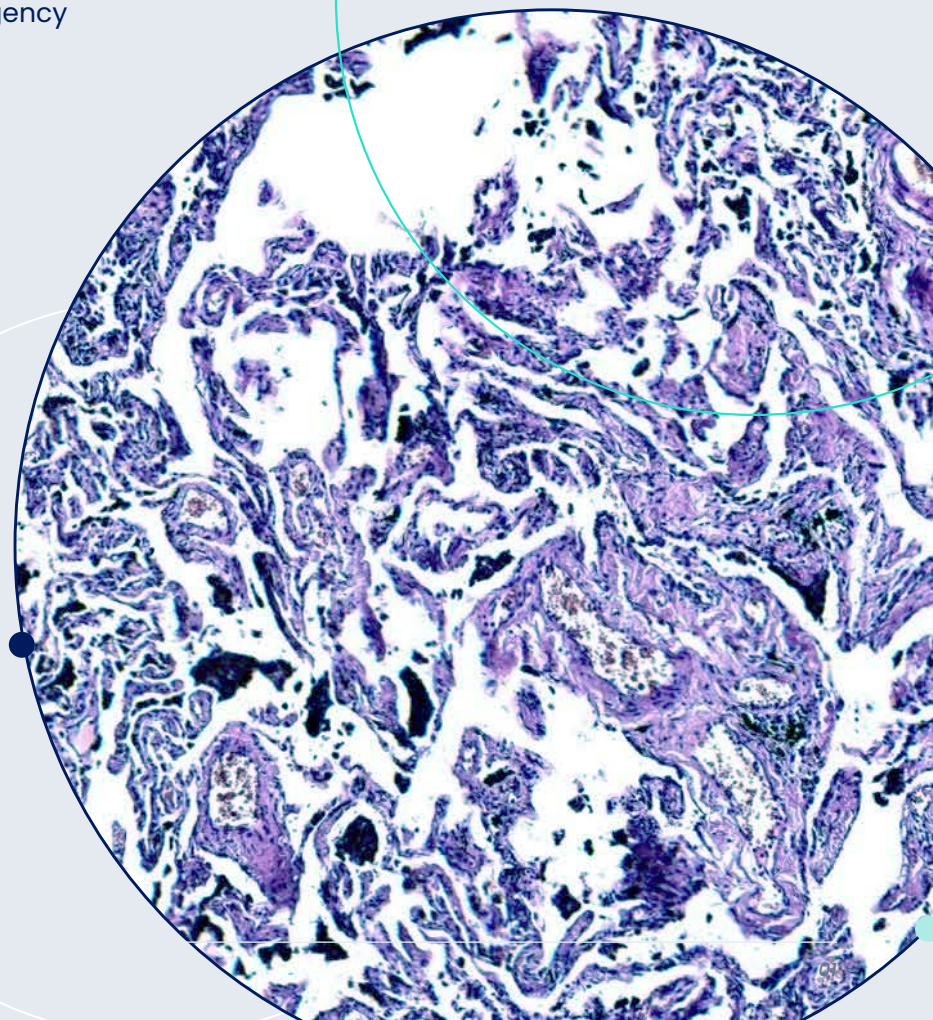
ARPA-H
Investor
Catalyst Hub
PART OF THE ARPANET-H
HEALTH INNOVATION NETWORK

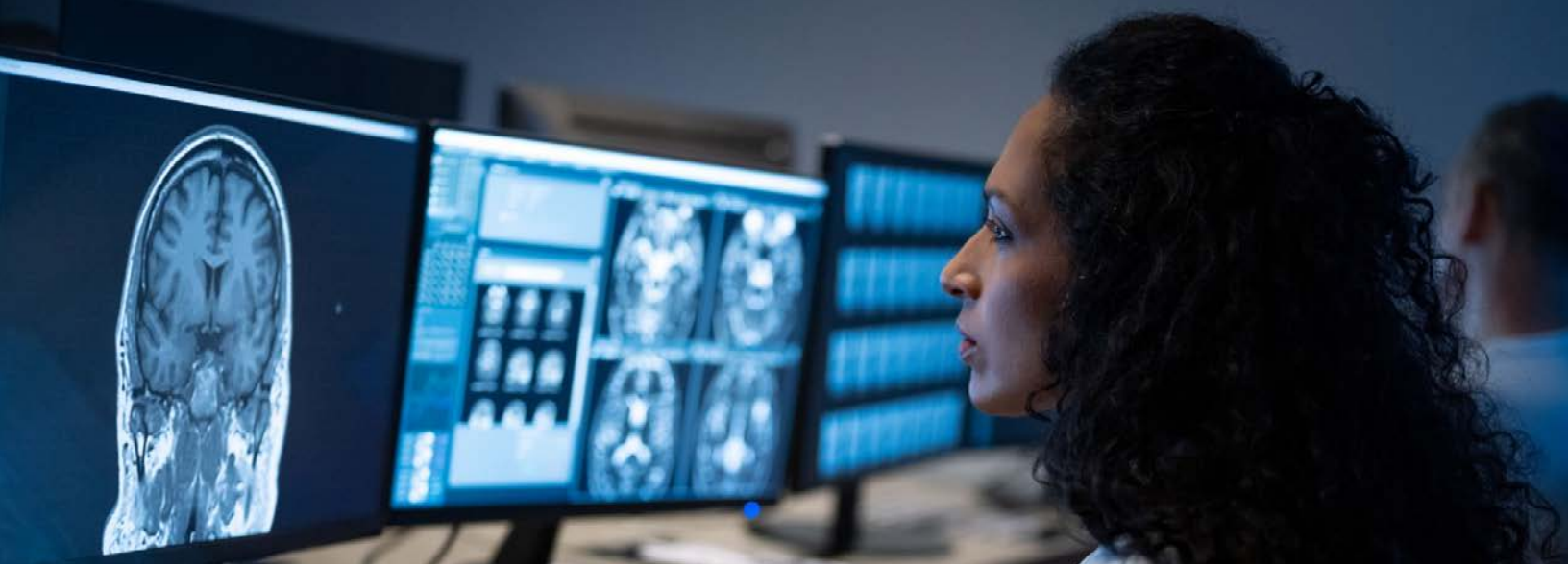
Medical Imaging Data Marketplace Survey Report

Authors: The Advanced Research Projects Agency for Health (ARPA-H)'s Investor Catalyst Hub aggregated and synthesized the results.



August 2024

Table of Contents



Introduction

Advancements in medical imaging data, artificial intelligence (AI), and machine learning (ML) have exponentially increased the versatility and usage of medical images. This progress has improved speed, reliability, and accessibility in modern diagnostics, treatment, and biomedical research.

An increase in the development of Software as a Medical Device (SaMD) incorporating AI and ML has resulted in new challenges for medical device and software developers. Limited access to high-quality, regulatory-ready medical imaging data for developing and testing new technology can be a significant barrier to innovation. Data quality deficiencies can also delay Food and Drug Administration (FDA) clearance or pre-market authorization, ultimately limiting the performance and availability of AI-enabled software and medical devices.

To address these challenges, the Advanced Research Projects Agency for Health (ARPA-H) and the Center for Devices and Radiological Health (CDRH) of the FDA seek to develop a medical imaging data marketplace (MIDM) and announced the exploration of a new effort to streamline access to affordable, high-quality, regulatory-ready medical imaging data at scale. An MIDM will connect existing databases, marketplaces, and data providers to a trusted platform that researchers and customers can use to find and affordably access the data needed to develop and test new algorithms.

In support of this effort, the Investor Catalyst Hub administered a network survey to collect feedback on the specific needs and challenges that medical imaging software and product developers, users of AI and ML medical imaging products, and private and public organizations face with utilizing, managing, and producing data for product development and evaluation.

The network survey findings represented in this report will be used to inform the model and strategy used for an MIDM and prioritize the types of data it should support. It will also help to develop a viable economic model to ensure the long-term sustainability of the marketplace.

Market Overview

Overall

The medical imaging industry is in a mature growth stage, with steady technological advancements and incremental systemic innovation over multiple decades. As defined by the FDA, the term "medical imaging" refers to several different technologies used to view the human body to monitor, diagnose, and support in the treatment of medical conditions.[1] This relatively fast-paced industry is seeing a steady stream of innovation driven by rapid technological advancements, particularly in areas such as AI, ML, 3D imaging, and the demand for rapid, point-of-care disease diagnosis.

United States

The United States contributes the largest share of the diagnostic imaging market, which can be attributed to ongoing technological innovation and an increase in the number of diagnostic procedures each year. U.S. health care spending increased from $2.6 trillion in 2010 to $4.3 trillion in 2021 and is projected to grow, on average, by 5.4% per year over the next decade.[2]

5.4%

projected compound annual growth rate (CAGR) in U.S. health care spending

Nominal spending on imaging increased 35.9% between 2010 and 2021, but as a share of total health care spending fell from 10.5% to 8.9%.[2] The total number of imaging examinations performed during this window in the employer-sponsored insurance population increased from 143.56 million to 146.81 million.[2] This growth in imaging spend was related to overall increased usage of the technology and the shift to using advanced imaging modalities.

Software-Specific

The market for AI in medical imaging is slated to reach $14.2 billion by 2032 (up from $762 million in 2022).[3] North America's AI in medical imaging market size accounted for $379.11 million in 2022 and is projected to reach around $5,680.48 million by 2032, expanding at a healthy compound annual growth rate (CAGR) of 33.60% from 2023 to 2032.[3] Ongoing innovation in imaging technology, such as computer-aided diagnosis (CAD), is expected to increase demand for these tools. The adoption of AI in medical imaging has catalyzed market trends in recent years and is expected to impact future growth positively.

End Users

In 2023, hospitals accounted for the largest market share of data imaging users. Nominal spending for medical imaging completed in the United States increased 35.9% from 2010-2021.[2] Rising demand for advanced imaging modalities and the integration of surgical suites with imaging technologies are some of the factors driving the segment growth.

Significant growth is expected in the diagnostic imaging centers segment during the next 10 years, owing to an increase in awareness about chronic diseases such as cancer and neurological diseases, as well as a nationwide push to outpatient acute care centers. There are 18,861 imaging centers across the U.S., according to Definitive Healthcare in February 2023.[4] The increased adoption of advanced technology, improved infrastructure, and high funding for the development of these centers is supplementing the segmental growth.

$14.2B

expected market for AI in medical imaging.[3]

18,861

imaging centers across the U.S.[4]

35.9%

increase in nominal spending for medical imaging from 2010-2021.[2]

Products

Products in the medical imaging marketplace include ultrasound, mammography, X-ray imaging, magnetic resonance imaging (MRI), digital pathology, and other scanning methods. Use of computed tomography (CT), MRI, and ultrasound in the United States increased rapidly from 2000 to 2006 and growth continues well into the 2020s.[5]

The rising rates of breast cancer and an increased demand for improved diagnostic solutions are driving the mammography market growth. Statistics from the World Health Organization report 2.3 million women diagnosed with breast cancer and 685,000 deaths globally in 2020.[6] The number of cases of breast cancer is generally higher in high-income countries, but women have a much greater risk of dying from this disease in low- and middle-income countries due to late diagnosis and limited access to treatment and care.[7] Currently, breast cancer diagnostic programs have been recognized widely in at least 22 countries.[8] Increased access to breast cancer screening systems and growing government initiatives to support clinical interpretation will likely lead to continued market growth.

As reported by Forbes, big growth is expected in the global digital pathology market. The market was valued at $740.26 million in 2021 and is projected to grow to $1,738.82 million by the end of 2028. That anticipated growth represents a 13.8% CAGR increase between now and 2028.[9] Health care players are increasingly focused on adopting IT infrastructure that enables the growth of precision diagnostic tools to address the rising prevalence of cancer and chronic diseases.[10] The use of AI in digital pathology has catalyzed the innovation funnel in this space and driven an increase in mergers and acquisitions. The overall high cost of operating these technologically advanced systems continues to act as a market inhibitor globally.[11]

Global Digital Pathology Market Growth

13.8%
increase
in CAGR by 2028

Underlying Market Factors

Tailwinds

The United States is experiencing growth in health care expenditure due to rising rates of chronic diseases and an aging population. In recent decades, the proportion of people older than 65 has significantly increased from less than 9% in 1960 to a forecasted growth from 17.3% in 2019 to 26.7% by 2050.[12] U.S. health care spending increased from $2.6 trillion in 2010 to $4.3 trillion in 2021 and is projected to grow, on average, by 5.4% per year over the next decade.[2] The Affordable Care Act increased access to preventive services—many involving imaging—for employer-sponsored insurance (ESI) beneficiaries.[2]

Increasing prevalence of lifestyle diseases such as heart disease, stroke, and diabetes, paired with rising interest in early detection tools, are fueling the demand for diagnostic imaging devices. Among Medicare beneficiaries, the increase in utilization was higher for medical imaging than other physician-provided services.[5] Within the diagnostic imaging space, there is a growing need for remote diagnostic technologies and point-of-care testing devices to improve access for rural and underserved populations.

This expansion of access to modern medical technology is fueling an increased focus on building advanced infrastructure that can support the use of AI in health care. Steep increases in imaging can be attributed to technical improvements, physician and patient demand, and strong financial incentives.[5] This increased investment and attention to insurance reimbursement has positioned the medical imaging industry, specifically imaging-based diagnostics, in a high market growth space.

Headwinds

The overall cost associated with medical imaging procedures may impact the opportunity for growth adoption, particularly in developing nations. From 2010 to 2021, nominal prices of hospital and physician imaging services increased by 22.87%.[2] In the United States, it is calculated that the average cost of an MRI, CT scan, and X-ray procedure range from $200-$2,200; $50-$1,500; and $50-$450, respectively, based on several factors including insurance coverage and location.[13]

The costs of these procedures are driven by the high cost of imaging equipment, increased data security infrastructure, and ongoing staffing and regulatory hurdles. Despite a global trend in the adoption of computer-assisted diagnostic imaging, the lack of standard and accessible imaging data for training of machine learning algorithms remains a barrier to progress in machine learning in medical imaging. To address these gaps, more effective methods are needed for data collection, de-identification, and management of images for research that use findable, accessible, interoperable, and reusable (FAIR) principles for scientific data management and stewardship.[14]

Network Survey Methodology

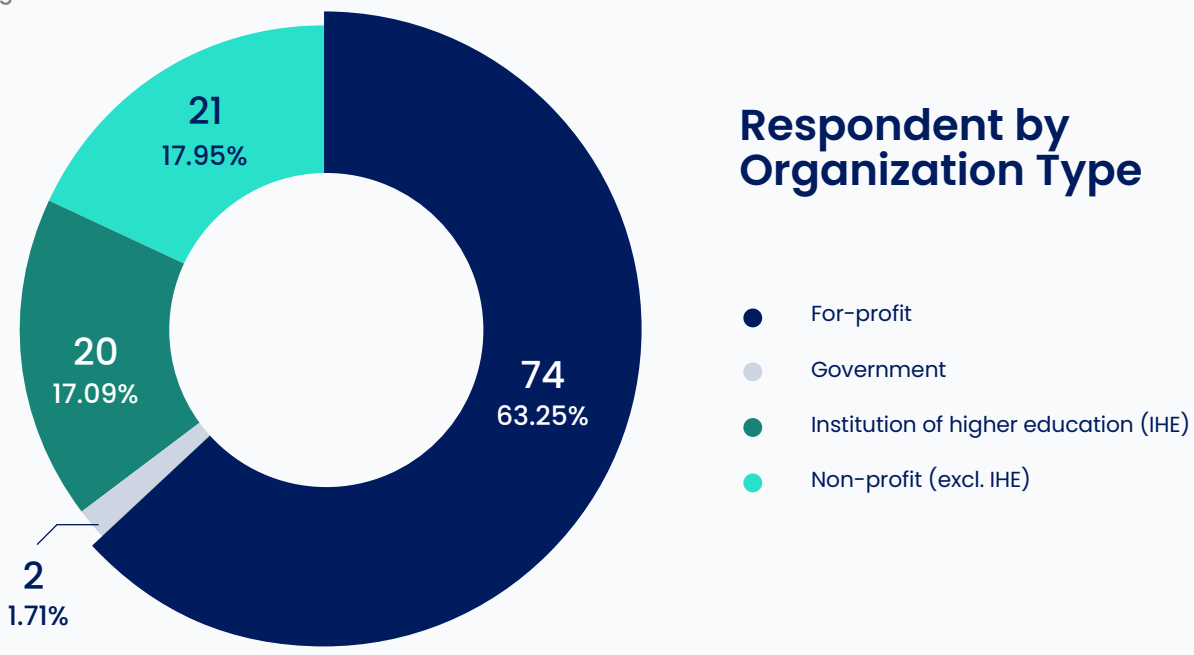
The MIDM network survey collected responses from institutions, researchers, and companies across the medical imaging ecosystem. A sampling method of relevant industry partners was used to distribute the online survey via the Investor Catalyst Hub and was open for one month. The survey outreach was targeted to data managers, data users, and data providers. An industry sampling was chosen to quickly gain broad preliminary insights into the ecosystem's operation and the needs of participants. The survey collected a total of 117 responses from organizations within the medical imaging and health data exchange ecosystem.

The Investor Catalyst Hub conducted a descriptive analysis (frequencies, counts, percentages, means) with the responses to single- and multiple-choice questions, and conducted a thematic analysis with the textual data responses to the open-ended questions. Where applicable, the hub conducted a frequency count on the major themes.

Survey Respondents

A total of 117 organizations within the medical imaging and health data exchange ecosystem responded to the network survey. The respondents represent for-profit companies, institutions of higher education (IHE), government institutions, and non-profit companies working across a number of primary sectors.
See Figures 1-2 for summary of respondents.

Figure 1

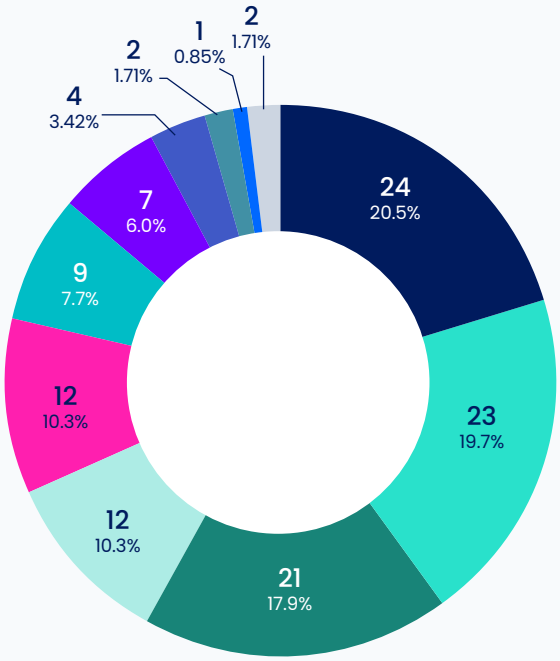


This is a multi-select question

Figure 2

Respondent by Primary Sector

- Medical device
- Research organization
- Health care system/organization
- Digital health
- Network/association
- Manufacturer
- Consultancy
- Pharmaceutical
- Clinical or contract research organization (CRO)
- Biotechnology
- Incubator/accelerator



This is a multi-select question

The survey asked respondents to self-select as one or multiple respondent types. Respondents were given data user, data manager, and data provider subject area-related questions based on their respondent type selection.

Medical Imaging Data Users

Understand needs to develop algorithms, obtain data, and meet data quality and regulatory standards.

- Radiology AI researchers and developers
- Digital pathology AI researchers
- Academic researchers
- Startups
- Medical imaging companies
- Medical imaging device manufacturers

Medical Imaging Data Managers

Understand where datasets are obtained, challenges obtaining and providing data, and operating models.

- Platform managers
- Data commons
- Databases
- Brokers
- Aggregators

Medical Imaging Data Providers

Understand the economic value of providing data and current challenges with obtaining and sharing data.

- Hospital systems
- Radiology departments
- Pathology departments
- Contract research organizations (CROs)
- Relevant government entities
- Medical imaging device manufacturers

The network survey included questions based on the respondent type selection in the following topic areas:

Data Needs and Usage

- Required services, standards, and capabilities the available data within the market would need to possess
- Specific AI and ML requirements and considerations

Economic Opportunity

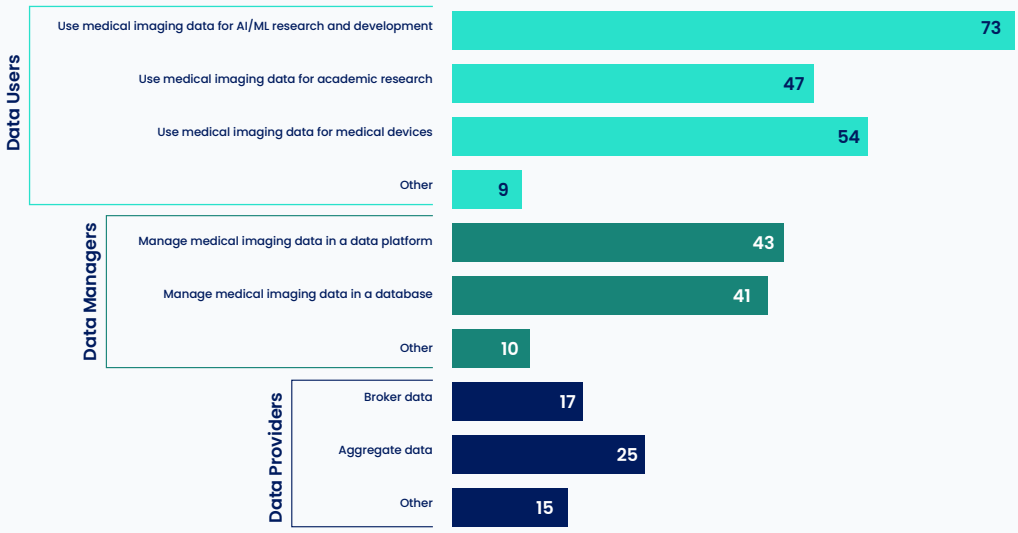
- Affordability requirements
- Data licensing and usage
- Incentive structures

Platform Requirements

- Required compliance, security, regulatory, and access needs
- Technical infrastructure preferences and requirements

Figure 3

Respondents by Self-Selected Type



This is a multi-select question

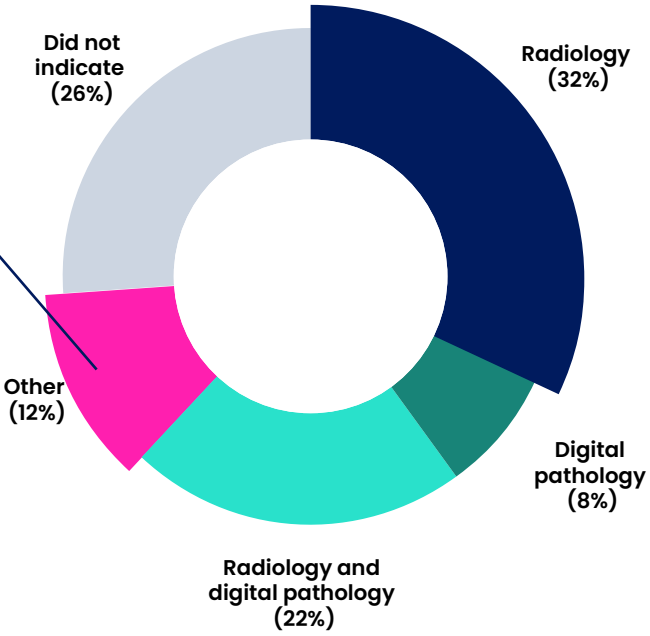
Data Insights by User

Survey respondents highlighted interconnectedness in the medical imaging data ecosystem by self-identifying as more than one role: 59% of respondents identified as a single role, 23% identified as two roles, and 18% identified as all three. The overlap in this ecosystem highlights that some organizations own their data experience from collection to reporting, while others have to work within the boundaries created by these data superusers.

Data users were asked which type of medical imaging data they needed: radiology, pathology, or 'other.'

Figure 4

Medical Imaging Data Needed



Other medical imaging data needed:

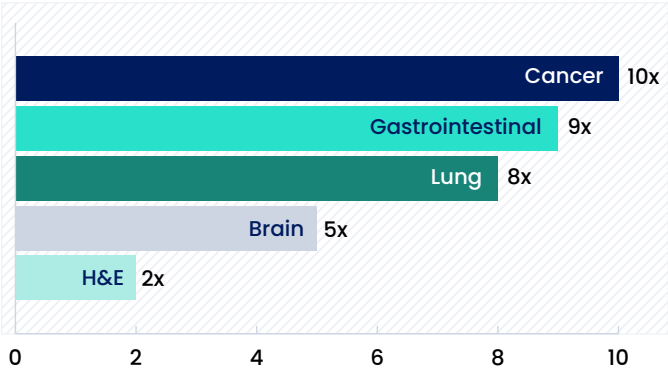
- Cardiology imaging (e.g., echocardiology)
- Ophthalomogy
- Product research and development (R&D)
- Metadata
- Dermatology
- Diagnostics/clinical data
- ENT (otoscopy)
- Endoscopy
- AI/ML training data

The following are multi-select questions

What data is needed in addition to digital pathology slides for breast and prostate cancer?

Pathology data users indicated:

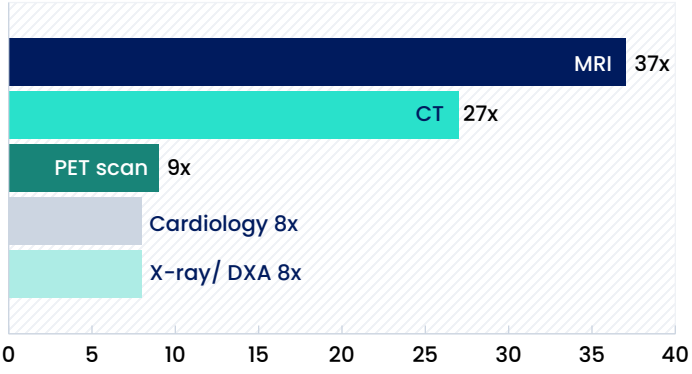
Figure 5



What data is needed in addition to mammography images?

Radiology data users indicated:

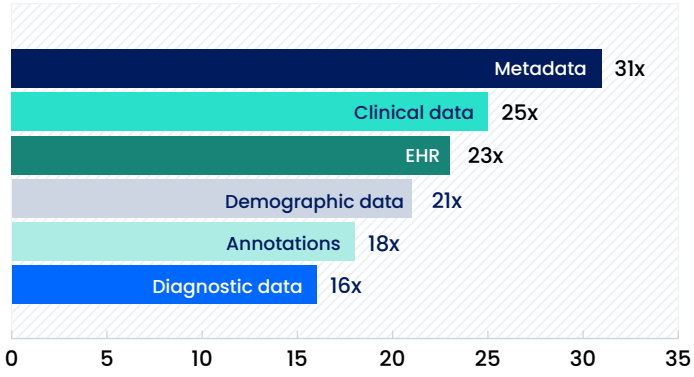
Figure 6



What do you accompany your data with?

Data managers indicated:

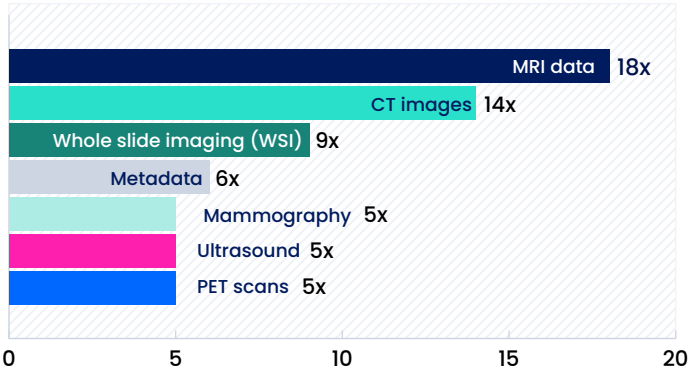
Figure 7



What data do you make available to data users?

Data managers indicated they provide:

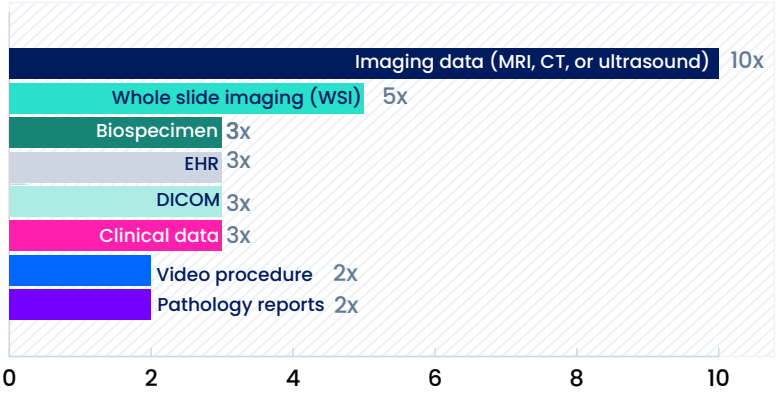
Figure 8



What do you provide researchers and data managers with?

Data providers indicated:

Figure 9



While responses indicate similar viewpoints among data users, data managers, and data providers, variations appear based on the size and tenure of the organization. For example, startups who participated in the survey highlighted decreased access to data, higher prices per image, and difficulty scaling data access due to variations across contracts and data usage agreements. In contrast, respondents from larger health care systems had access to data but noted the types of data they could access influenced the types of projects they were able to conduct.

Survey responses show a need for a variety of radiological imaging data, and highlighted specific modalities like MRI and CT that would be especially impactful. However, within these modalities there were a wide range of conditions and targeted organs needed by data users. Pathology data users likewise indicated a wide range of data needed. Based on the responses, it is difficult to clearly identify a specific third use case beyond mammography and prostate/breast WSI. While not surprising, this illustrates the diversity of research and applications of AI/ML to radiology and pathology and the acute need to reduce barriers to accessing medical imaging data.

Responses from underrepresented data providers were limited. Six respondents identified as a socially or economically disadvantaged business (8a certification), Small Disadvantaged Business (SDB), Historically Underutilized Business (HUBZone location), or a Veteran-Owned Small Business (VOSB). Fourteen identified as minority-owned, women-owned, B Corp, or LGBTQ-owned. Some respondents indicated that they have engaged with underrepresented data providers. These groups represent an important gap that may be missing from the survey and is an area that is recommended for further exploration.

Survey Findings

MIDM Foundational Considerations

Compliance Framework

Survey respondents used a wide range of security, interoperability, and regulatory standards depending on their specific use cases and applications. Many of these standards are, however, only relevant to cloud service providers or applications that store and use protected health information (PHI). Data managers frequently noted that they de-identified their data to maintain compliance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. While not explicitly stated by all respondents, it is believed that this is primarily done using the Safe Harbor method rather than the Expert Determination method.

Data managers' most common security standards and certifications:

- HITRUST certification
- ISO 27001
- ISO 27018
- ISO 27701
- SOC 2 Type 1
- SOC 2 Type 2
- NIST 800-53

Data managers' most common data interoperability and transfer standards:

- FHIR
- HL7v2
- DICOM (DIMSE Protocol or DICOM Web Services)

Data managers' most common regulatory compliance standards:

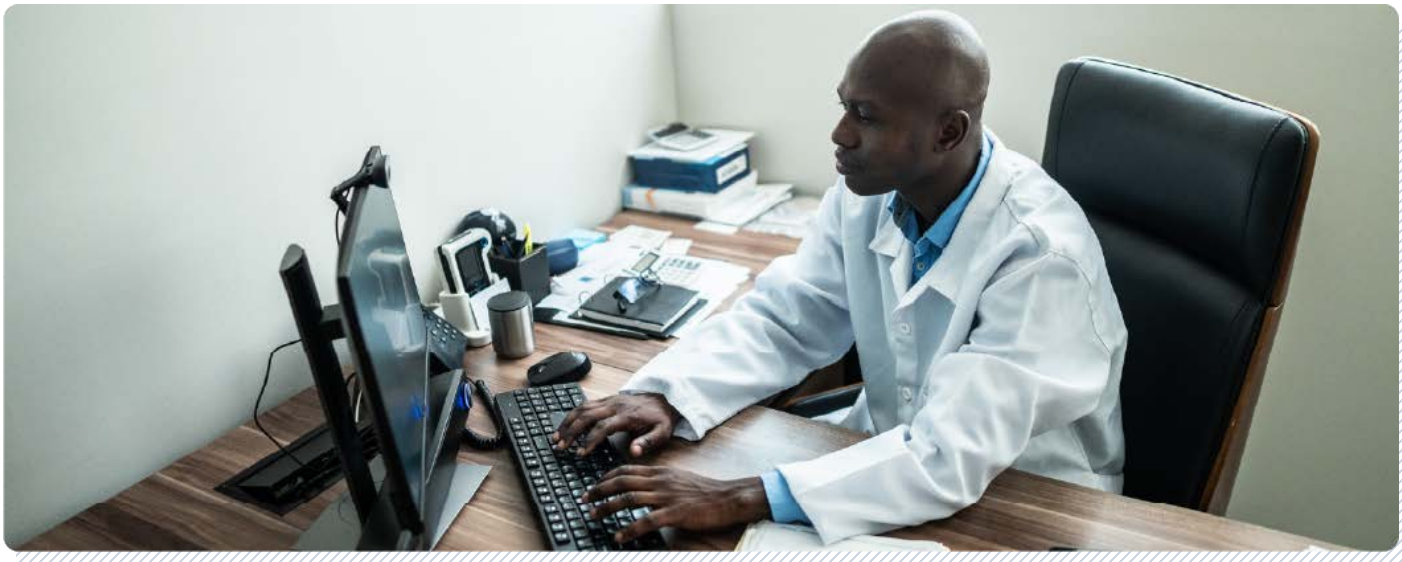
- HIPAA
- 21 CFR Part 11
- GDPR
- PIPEDA

Data users highlighted some foundational minimum dataset requirements. Respondents indicated de-identification as crucial to meeting HIPAA and General Data Protection Regulation (GDPR) anonymity requirements. Data providers outlined the need to de-identify their data by removing burned-in pixels from images, personally identifiable information (PII) from the image metadata, and any PII in associated pathology or radiology reports

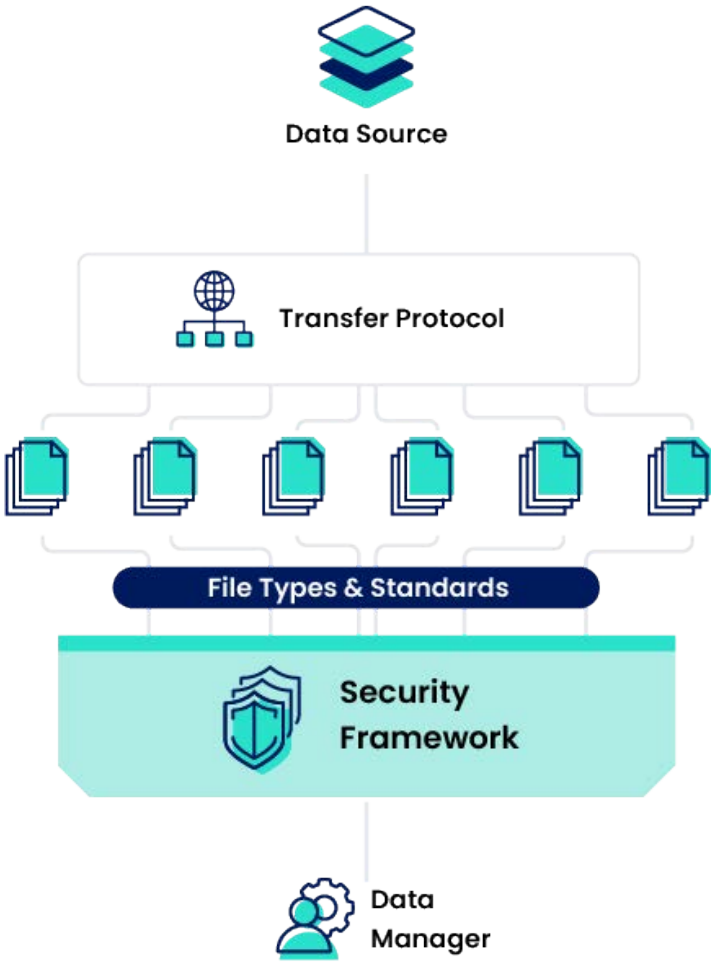
However, the de-identification process frequently removes demographic information that directly impacts data users' ability to build representative datasets and unbiased AI models. For example, participants noted many databases redact data beyond the Safe Harbor de-identification requirements, stripping information about ethnicity and race. Missing demographic information can be filled in using privacy-protecting record linkages (PPRLs), but this is a relatively new advancement and is not standard in all datasets.

In addition to needing demographic information, data users also need access to the pathology and radiology reports associated with the medical image, but only 21 out of 40 data providers said that they provide them to data users.

All three types of respondents highlighted the importance of interoperability, indicating that an MIDM will need to support more than one set of standards and certifications to enable broad adoption of the marketplace. DICOM was the primary standard mentioned by data providers, but they also indicated the Neuroimaging Informatics Technology Initiative (NIfTI), Tag Image File Format (TIFF), and stereolithography (STL) as commonly used standards. DICOM is most prevalent in radiology, but respondents noted substantial variation even within the DICOM standard depending on the imaging device used. Pathology data managers noted needing to support more than a dozen formats, with some file types unique to the scanning equipment used.



Data managers indicated their transfer protocols and extraction tools are enabled by using the DICOM standard, most commonly executed using a shared cloud platform, secure file transfer protocols (SFTP), or Fast Healthcare Interoperability Resources (FHIR). However, roughly an equal number of data managers reported not allowing data transfers. Respondents note an MIDM will need to focus on a set of acceptable security frameworks that seamlessly navigate the challenges of varying file types and standards.



Product Infrastructure

Centralized Versus Federated Architecture

According to data providers, hospitals find a federated network more comfortable for storing and maintaining their data. However, federated networks can increase the difficulty for data providers in integrating with a data sharing platform and substantially increase the challenge of providing access to data users, especially when the data they require is spread between multiple data providers in different cloud environments.

A centralized architecture for de-identified data eases the challenges of accessing the data, but increases storage costs and affords data providers less control. As a result, many of the data providers that responded to infrastructure questions choose to host data using a hybrid approach. This was reflected among a class of data managers that found success aggregating data from multiple providers, then working as an intermediary to make it available to data users. For this hybrid approach, de-identified data was shared with the manager when it was ready to be made available to the data users. In contrast, data managers that connected users to providers but did not act as an intermediary used a decentralized approach rather than a hybrid one.

Respondents highlighted numerous challenges with decentralized approaches, including that the bandwidth of a hospital is a rate limiter. Hospitals may lack the resources to manage a federated interface over time, may not allow case images to leave their network, may change picture archiving and communication system (PACS) vendors, be impacted by the cost of on-premise de-identification, or may just have slow turnaround times for indexing. As noted by one respondent working with their hospital data providers, to upload a petabyte scale dataset (about 1 million studies or the amount needed for a foundation model), at 5GB per hospital per day, even with 20 hospitals in parallel, could take 10,000 days—more than 27 years.

> *As noted by one respondent working with their hospital data providers, to upload a petabyte scale dataset (about 1 million studies or the amount needed for a foundation model), at* ***5GB per hospital per day, even with 20 hospitals in parallel, could take 10,000 days—more than 27 years.***

Furthermore, respondents identified significant usability issues with a fully federated learning approach. If data users are only able to run their model without directly interacting with the data, it presents significant hurdles to debugging models, handling exceptions due to variations in data standardization, and the inability to label or annotate the data for training. Respondents that identify as both data users and data providers noted that in order to be successful, the marketplace model will likely need to incorporate a distributed framework for data storage.

Data Characterization and Search

Survey responses indicated three main R&D use cases for the marketplace: search and query, classification, and AI models. Organizing the marketplace by anticipated use could ensure a better user experience.

1. **Search and query:** Finding cases and images to assess the research question
2. **Classification:** Establishing the necessary metadata, annotation, and outcomes to enable analysis and categorization of data
3. **AI model:** Determining and finding the necessary data to train the AI model, performance metrics (e.g., sensitivity/specificity, comparison to current standards) and outcomes of the output of their models

Search and Query

Data users say the ability to search and query data is incredibly important for discovering datasets and determining the viability of projects. Filter, search data, or query was mentioned by numerous respondents, indicating the importance of supporting keyword-based search, filtering by metadata, and advanced query capabilities. Data users also need to be able to search for data that fills demographic gaps to create representative datasets.

Survey respondents indicated the infrastructure of an MIDM will then need to support two features: the ability to search for available data before purchasing and the ability to access that data. Prioritizing a minimum requirement for metadata will be helpful for effective search of available data.

Aggregation and Classification of Data

Classification of data–including patient demographics, image modality, anatomical site, and diagnoses–is a fundamental process to standardize metadata, ensuring accurate search and retrieval. This process consists of collation, curation, and annotation. Survey respondents brought up classification 17 times in their responses, and indicated marketplace adoption would be improved by offering flexible architecture that would enable users to continue to utilize their own processes.

According to survey responses, the marketplace can prioritize supporting data according to potential for overall high impact, disease area, current availability of data sources or data types, and/or focusing on expanding to specific intended uses.

Respondents prioritize disease data by:

Incidence/volume, likelihood of impact on outcome, difficult detection with current methods, high risk/aggressive, rare disease experience may be lacking and/or experts are more likely to miss, potential use by non-experts at proof-of-concept phase

Respondents' highest-ranked data by type of disease:

Cancer/solid tumors, Alzheimer's disease, Parkinson's disease, coronary heart disease, abdominal aortic aneurysm (AAA) progression, renal disease, diabetic retinopathy, metabolic diseases (e.g., evidence of fatty liver disease/metabolic dysfunction-associated steatohepatitis)

Respondents' highest-ranked data by intended purpose:

Detection, progression, treatment decision, risk determination, treatment monitoring, post-treatment monitoring, clinical trial monitoring

AI Models

Survey respondents indicated they are utilizing imaging data and associated metadata to create AI-enabled tools using a wide range of techniques; their data needs ranged from 100s to 100,000s. About half of the data users that answered the question said they were using fewer than 1,000 images. This was noted most often for fine tuning or initial experimentation, not the development of a robust application.

Based on responses and public data sources, +10,000 samples would likely be needed for any FDA filings. One can also extrapolate from other survey questions that the reported average dataset sizes reflect pervasive challenges in collecting and annotating data, restricting the range of machine learning techniques that can be applied.

Data managers and data providers noted the importance of multimodal data and the creation of AI/ML-enabled solutions as a continually important, growing trend. Survey responses from these respondents indicated a marketplace will be most successful if data is multimodal with robust metadata available to ensure the right data is being utilized for each use case.

Multimodal data was mentioned twenty-one times by survey respondents. As suggested by many data managers, multimodal data could be supported by incorporating PPRLs.

One respondent noted, "A useful data repository for AI/ML in medical imaging in the U.S. can have comprehensive metadata, be easily accessible, affordable, and sustainable, ensure data quality and representativeness, include diverse patient populations, and support user-friendly querying and data retrieval mechanisms. The dataset must apply FAIR data principles in which the data are Findable, Interoperable, Retrievable, and Accessible."

Inconsistency across data users suggests there is no one-size-fits-all approach or methodology for AI. The amount of data required to create an initial AI/ML-enabled proof of concept algorithm compared to the amount of data for a regulated product varies greatly. Survey responses show the number of cases, rather than the number of images, is a better measure to see how much data is needed per algorithm.

> *"A useful data repository for AI/ML in medical imaging in the U.S. can have comprehensive metadata, be easily accessible, affordable, and sustainable, ensure data quality and representativeness, include diverse patient populations, and support user-friendly querying and data retrieval mechanisms."*

Survey respondents noted that to speed up AI/ML-related submissions, the marketplace will need to have large enough datasets available that will allow for a full FDA submission. As noted by data users, data representation majorly impacts the ability to generate valid models. Respondents suggested a mitigation involving combining data across organizations to create a complete dataset.

Figure 10

AI/ML Intended Uses by Respondents

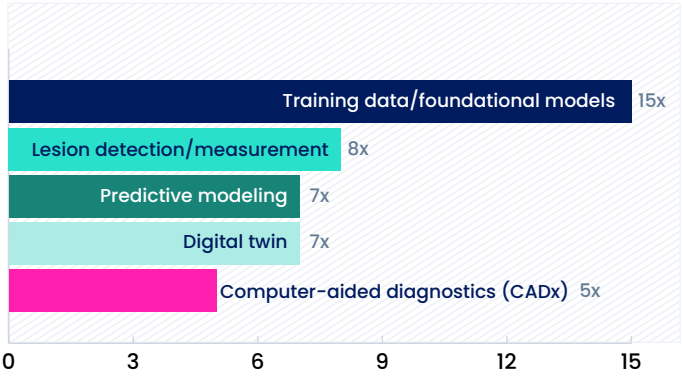
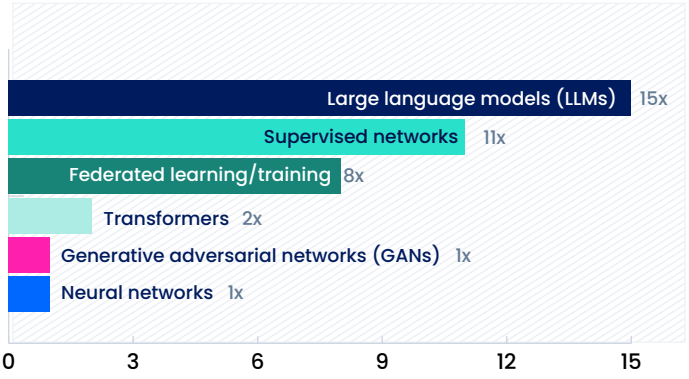


Figure 11

AI/ML Methods by Respondents



Interoperability

Interoperability standards varied among respondents based on the maturity of an organization, the breadth of products they create, and the intended use of those products. DICOM is the most established standard for images—mentioned by almost all respondents. DICOM was mentioned 145 times throughout survey responses. Seventeen respondents indicated they use DICOM for standardization. Ten explicitly mentioned conversion to other standard formats: NIfTI, Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine (SNOMED), Medical Subject Headings (MeSH), Health Level Seven (HL7), and FHIR.

More than two-thirds of data users and data managers noted the use of in-house tools to support their data work. Survey respondents also frequently indicated the essential need for new tools to integrate with their IT infrastructure, including custom-built software (especially for data annotation). Integration could be managed with data downloads or data licensing, or alternative cloud-based Application Programming Interface (API) integrations.

PPRL tools and tokenization technology were also mentioned across all three data user types. One respondent noted that to integrate multimodal and external data sources, PPRLs are needed "to ensure two sources have information with regard to the same patient. Additionally there will be a need for confirming images have shared anatomy and/or the same image over time. The marketplace can consider having these tools to support the ability to identify and de-duplicate datasets."

Market Consideration

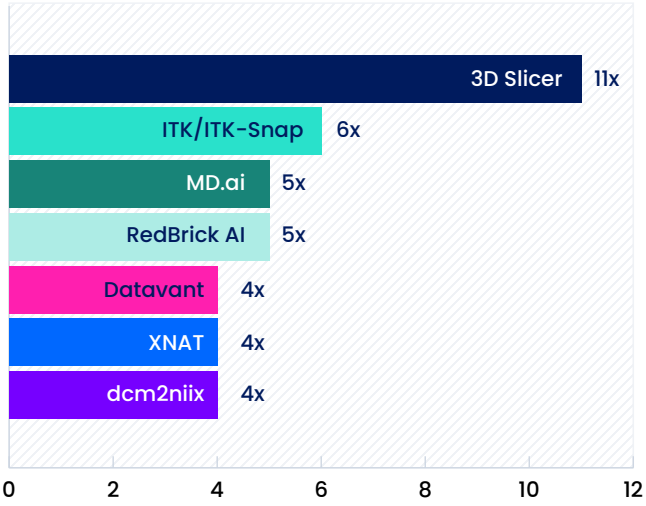
Due to the broad range of organizations the respondents represent, the survey collected a variety of responses across data users, data managers, and data providers. To serve the needs of these diverse perspectives, the marketplace could present a menu of offerings with certain features, datasets, and services available at no cost, low cost, or a la carte to allow for smaller and less established organizations and public researchers to engage with an MIDM.

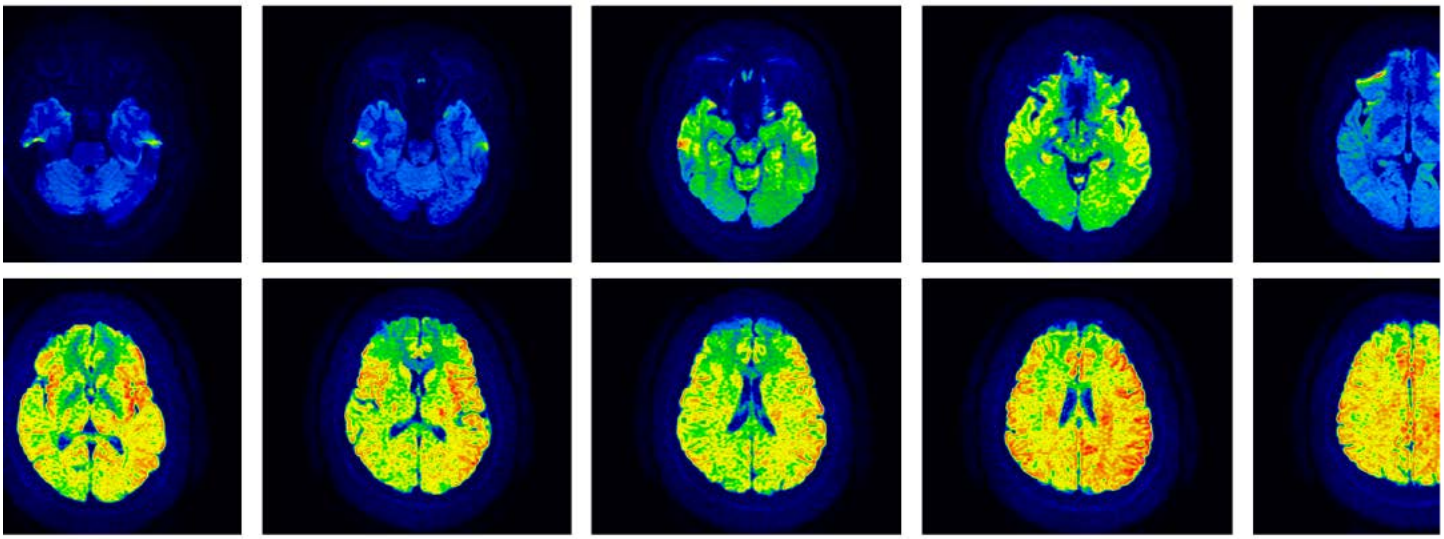
Respondents recommended flexible pricing options for data–either pay-per-image or umbrella agreements–clear pricing when additional EHR or clinical data are available, negotiable pricing for annotation and labeling services, and free access to search features. Data users universally requested transparent, public pricing that does not require bespoke negotiations for each purchase. Offering limited free data access on the platform may also increase use and adoption of the tool, and could improve access for early-stage innovation initiatives.

Data managers and data providers noted that a marketplace will need to offer or integrate with a set of tools that allows users to inspect data, and recommended considering additional annotation and regulatory services. Third-party tools are frequently used across data users.

Figure 12

Top Third-Party Tools Mentioned in Survey Responses





Challenges Facing AI Research and Productization

Challenges

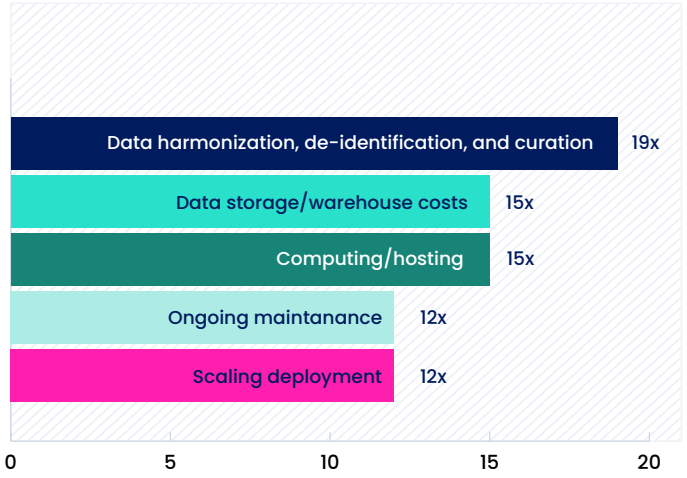
Survey respondents highlighted a few areas where they have experienced inefficiency throughout their medical imaging data use experience: delays, cost, and contracting. Delays were mentioned 39 times by respondents.

High costs and delays in gaining access to necessary data is a pain point for survey respondents.

- Costs or funding are the biggest challenge in finding or obtaining medical imaging data (mentioned 12x by data users)
- Cost is the greatest challenge faced when sharing data (mentioned 7x by data managers)
- Cost must be overcome to participate in an MIDM (mentioned 8x by data providers)

Figure 13

Top Cost Drivers Identified by Data Managers



Data users indicated delays in generating results ranging from

2 months to 2+ years.

Contracting Issues

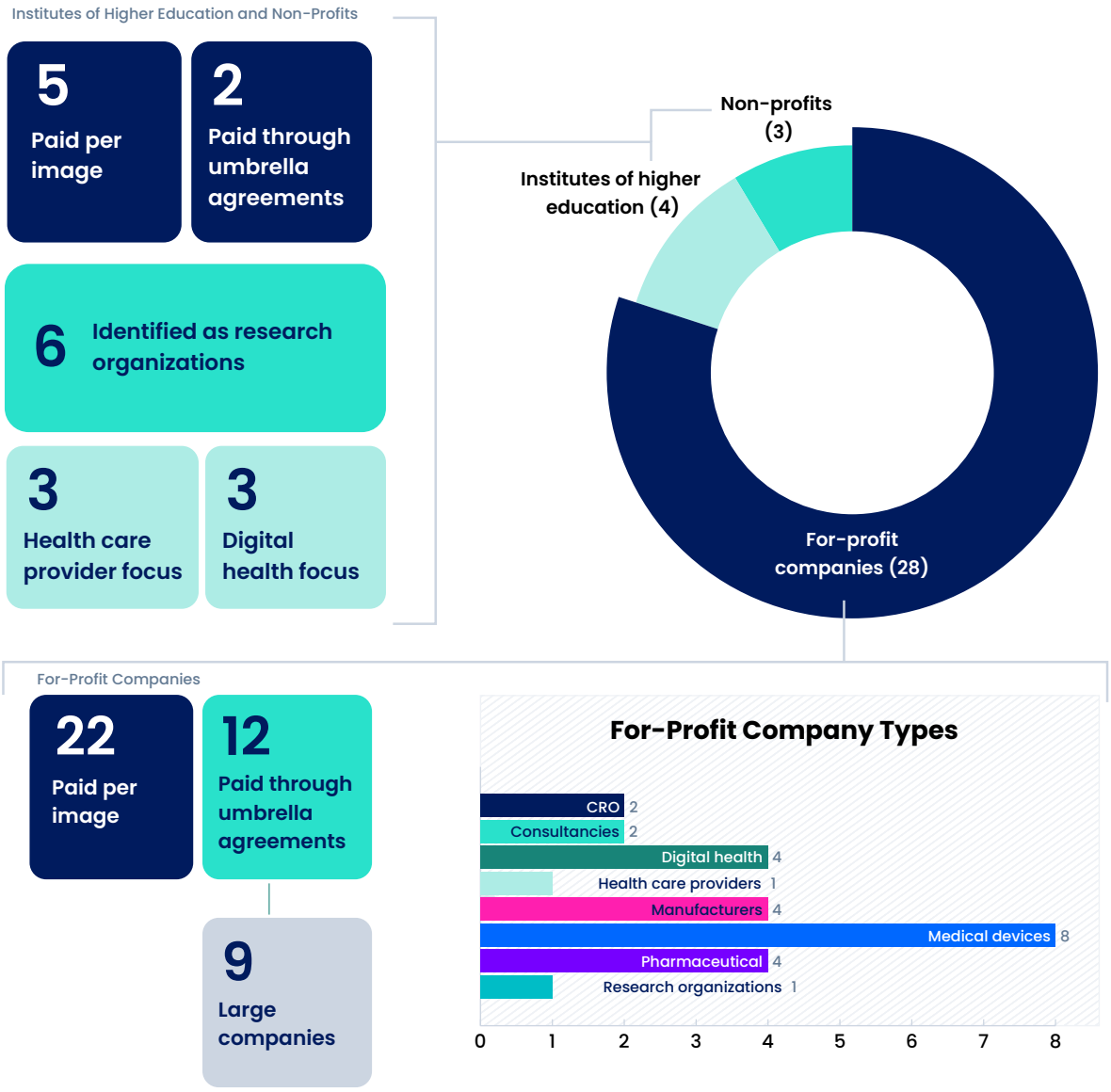
Survey respondents indicated the marketplace will need to support streamlining health system contracting, as well as to incentivize health care providers to share their data within the platform. Legal and privacy issues further complicate data sharing and access: Respondents noted that many centralized data collections are encumbered by data sharing restrictions which would limit the ability to provide that data to other users. Simplification of the contracting process will be a key initial requirement for an MIDM to be successful, including addressing providers and data users' fears around how the data will be used.

Of the data users, 35% receive their data directly from health care providers. Many of these health systems have unique contracting processes to access their data and/or offer their data within their own tools and systems. All three respondent groups mentioned a desire for upfront cost transparency in these contracts. Contracting was also frequently cited as a barrier to obtaining adequate medical imaging.

The majority of respondents paying for data through umbrella agreements identified as large for-profit companies, while smaller organizations more commonly pay per image. Of the respondents, 35 indicated they had paid for data.

Figure 14

Breakdown of Respondents Paying for Data



Data Quality and Completeness

General Data Quality Issues

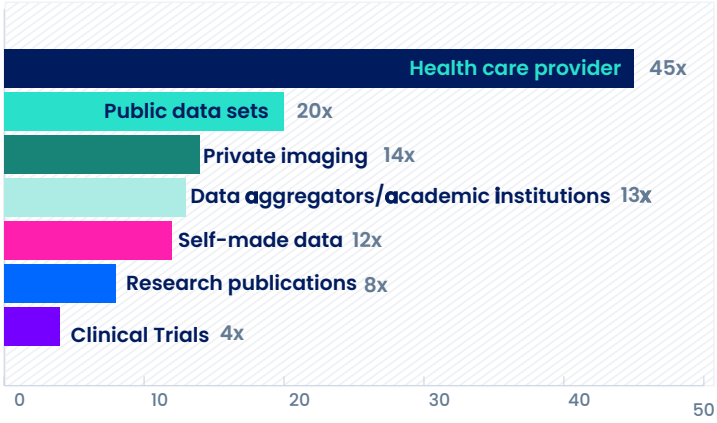
Survey respondents indicated an overall lack of trust that underlying images, metadata, and labeling/annotations will reach the standard and quality the researchers require for their projects. Extrapolating, this status quo could be addressed by upholding rigorous standards through an MIDM. Data users report filling in these gaps with manual work and software tools, and express a desire for implicit indicators within the marketplace showing which data meets which quality standards that align with which regulatory standards.

> ***Trust** was mentioned*
> **22x**
> *by respondents as a reason for not using available data.*
>
> *(e.g., trust of providers having the right metadata, trust of automated outputs of labels, trust of providers to share data in the right de-identified format.)*

Data users are collecting data from an array of different sources with different user experiences. *See Figure 15 for survey responses.*

Figure 15

Data Sources for Medical Imaging Data



Data curation is a major cost for both data users and data providers. Eighty-six percent of data providers said they curate or sometimes curate their data. Data managers noted that de-identification is an expensive bottleneck of the data harmonization and curation process, but it can be alleviated by the use of cloud-based technologies.

Data users indicated that they have unique sets of criteria to meet their own data quality requirements, citing this as a reason why public data sources may not be usable for their programs. Ten respondents specifically indicated that public data was not of high enough quality to be used.

Survey respondents indicated a few types of general data quality concerns: annotations, image data (e.g., metadata, pixel, and reports), interoperable formats, and other metadata (e.g., clinical, outcome). Underlying data quality issues around image quality and image metadata was the most common reason that public data sources were not adequate for data users' goals.

Data users cited image metadata, associated radiological and pathological reports, and image pixels as the top image quality issues.

Survey respondents across all three user types are utilizing different data transfer and standardization formats. For data managers, support for interoperability between various standards could allow more groups to participate in contributing data, consuming data, and providing services/tools. All respondent types noted that the more the marketplace can be designed to incorporate the current workflows and tools of researchers, the more it will be adopted at scale.

Metadata provided by data providers and data managers varied greatly based on type of institution and size of the organization with the associated amount of staff available to support metadata creation. Efficient metadata as noted by respondents supports better understanding, selection, and use of the data and could include clinical data and notes, lab or Rx data, and other EHR data. Data providers demonstrated a focus on what minimum viable dataset they can provide, including core data and metadata, with the goal to submit quality data that will attract multiple users. A marketplace will need to support the storing and processing of multiple types of imaging data and their associated metadata.

One data user noted the importance of metadata to support their analysis, stating, "Often, imaging data cannot leave the network of the health system that originated the image, but they can provide access to imaging data within the network to conduct analyses. Metadata in DICOM imaging files, as well as unstructured imaging impression reports, can usually leave the health system's network, but [must be] be tokenized by tokenization engines like Datavant or HealthVerity to make the imaging linkable to other clinical outcomes data (claims, EHR, labs, etc), as well as [social determinants of health] data. Often recruiting partners to provide data for certain rare outcomes is challenging, but even for more prevalent clinical domains it requires an effort to recruit health systems partners to contribute data."

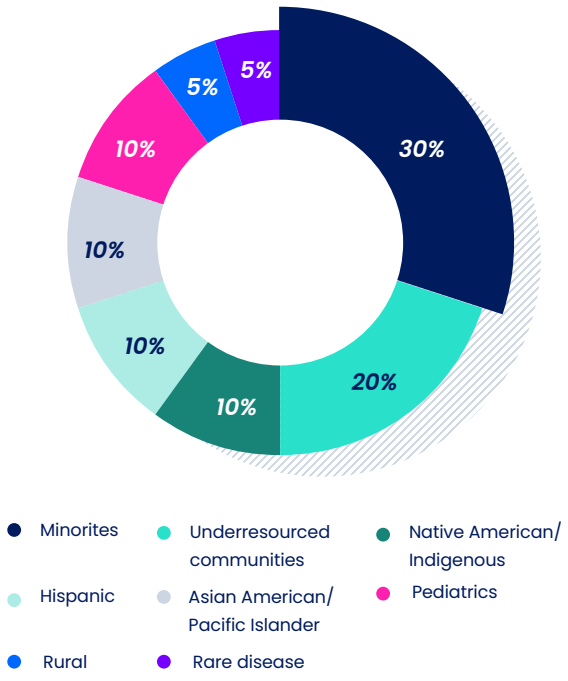
Data Representation Issues

Lack of data representation across demographics and disease areas serves as a market barrier as it can cause delays in development or total project abandonment. Representation was mentioned as a top reason for delay and project abandonment by respondents, with identified population gaps including diversity in racial and socioeconomic, geographic, rare disease, and age representation.

Fifty data users indicated they were able to have representative datasets, 24 indicated they were not, and 12 were not sure.

Figure 16

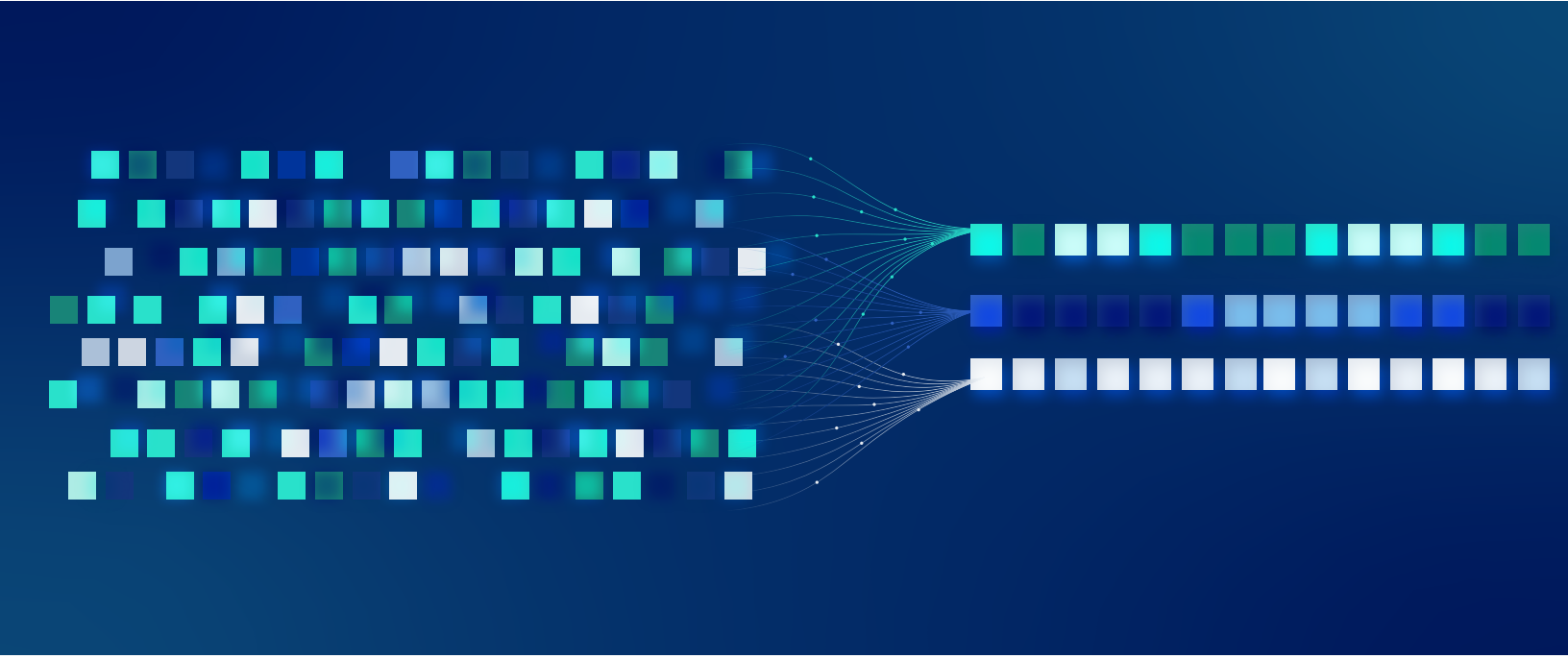
Population Gaps Identified by Respondents



Survey respondents indicated a willingness to pay to use data when that data meets their specific intended uses and fills data gaps. Of the 35 that purchased data, 12 were still unsure or unable to obtain representative data for their intended population. These respondents attempted to fill these gaps with data from multiple locations, taking all data for a given indication, and sought out diverse datasets, direct partnerships, and community engagement. This creative gap-filling highlights that even when capital is available to fill datasets, the data is still not available as the researchers need it.

Data users are seeking out data from free and paid sources to build representative datasets for their submissions. These users indicated there is a lot of work needed to get a broad network, inclusive of smaller providers working in underserved areas, prepared to engage and build sufficiently large datasets with enough representation and diversity to fill data gaps across different use cases. For example, when developing AI detection methods for finding cancer in digital mammograms, one respondent noted needing thousands of mammograms since cancer will only appear in four to five images out of 1,000 cases.

All three respondent groups noted that the marketplace could look to fill gaps for data from smaller and community-focused health care providers to broaden reach beyond larger research institutions and health systems. These smaller, community data managers and data providers (e.g., community hospitals, private clinics, and federally qualified health care centers) may need an incentive to participate because they are resource-constrained. Respondents working with these image providers recommended support approaches including: free data storage, revenue sharing agreements, free technical support for data sharing, and data de-identification or enrichment support.



Annotation Issues

Annotation of medical imaging data is a pain point for respondents and causes delays depending on the needs of data users at different stages of a project. Some users cited spending up to two hours per case. Currently, there is no single adopted standard for annotation methodologies, and data managers suggest aligning on how annotations can be applied, as well as who can apply them, if an MIDM will include annotated data. As noted by all three types of respondent groups, guidelines for the generation of annotations with the flexibility to apply appropriate methods for specific applications would allow for the continued application of in-house tools and processes that are commonly used today.

Data managers noted that medical images were often not annotated prior to AI or other analysis. Thirteen respondents indicated that annotating complete datasets can take from a few weeks to over a year, with most indicating at least six months required. Four respondents shared that it can take up to two hours to annotate a single case, indicating it could take one to two weeks to completely process a record (which respondents noted includes additional work on top of the annotation).

Respondents use in-house tools for annotation rather than relying on the quality of annotations received from data sources. According to survey feedback, widely adopted methods for annotation which meet regulatory-grade standards are needed across the imaging community, to enable streamlined and consistent annotation processes at the imaging site.

Currently, data managers and aggregators offer annotation and labeling services. Their responses demonstrate that longer-term services or automated algorithms that offer annotation services support the sustainability of the marketplace.

ARPA-H
Investor Catalyst Hub
PART OF THE ARPANET-H HEALTH INNOVATION NETWORK

Conclusions

The survey respondents provided significant insight into the needs of the medical imaging ecosystem. Ultimately, their feedback reinforced the need to convene a critical mass of stakeholders to tackle the pervasive issues preventing the distribution of medical imaging data and stymying the progress of the research community. Numerous perspectives were offered, often with competing visions for how to solve these problems, while highlighting a consistent set of needs and challenges.

These challenges often manifested in different ways for each respondent type, but stem from the same root causes. Standardization, interoperability, transparency, security, access, affordability, and trust all emerged as themes that must be addressed to establish a viable solution. Significant work must be undertaken to balance the concerns of data providers with the needs of data users, but numerous respondents highlighted models and approaches that have the potential to do so.

To meet the need for representative data a solution must actively break down the participation barriers that prevent data providers who serve minority and underrepresented populations, often with considerably lower resources, from sharing that data. In addition to data, it must provide tools for the research community to easily identify potential bias within their datasets to ensure that the next generation of AI and ML products work for all Americans, regardless of race, sex, age, or geographic location.

Overall, the survey outlined that there needs to be more than a simple marketplace for data–there needs to be an evolving exchange that cuts across silos, incorporates new participants, provides tools and services to researchers, promotes patient privacy and security, and furthers a common goal of improving quality, access, and affordability of healthcare.

Citations

1. Medical Imaging, U.S. Food & Drug Administration
2. Decomposition of medical imaging spending growth between 2010 and 2021 in the US employer–insured population Horný M, Chang D, Christensen EW, Rula EY, Duszak R Jr. Decomposition of medical imaging spending growth between 2010 and 2021 in the US employer-insured population. Health Aff Sch. 2024 Mar 27;2(3):qxae030. doi: 10.1093/haschl/qxae030. PMID: 38756926; PMCID: PMC10986240.
3. AI in Medical Imaging Market Expected to Increase to $14.2 Billion by 2032 Contreras, Biriana, Managed Healthcare
4. How many imaging centers are in the U.S.? Definitive Healthcare
5. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016 Smith-Bindman R, Kwan ML, Marlow EC, Theis MK, Bolch W, Cheng SY, Bowles EJA, Duncan JR, Greenlee RT, Kushi LH, Pole JD, Rahm AK, Stout NK, Weinmann S, Miglioretti DL. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. JAMA. 2019 Sep 3;322(9):843-856. doi: 10.1001/jama.2019.11456. PMID: 31479136; PMCID: PMC6724186.
6. Breast Cancer, World Health Organization
7. Improving access to medical imaging for more patients, GE Healthcare
8. Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review Hussain S, Mubeen I, Ullah N, Shah SSUD, Khan BA, Zahoor M, Ullah R, Khan FA, Sultan MA. Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. Biomed Res Int. 2022 Jun 6;2022:5164970. doi: 10.1155/2022/5164970. PMID: 35707373; PMCID: PMC9192206.
9. Market Factor Synergy Signals Exciting Growth for Digital Pathology, Dan Lambert, YEC Council Post
10. Digital Pathology: Transforming Diagnosis in the Digital Age Kiran N, Sapna F, Kiran F, Kumar D, Raja F, Shiwlani S, Paladini A, Sonam F, Bendari A, Perkash RS, Anjali F, Varrassi G. Digital Pathology: Transforming Diagnosis in the Digital Age. Cureus. 2023 Sep 3;15(9):e44620. doi: 10.7759/cureus.44620. PMID: 37799211; PMCID: PMC10547926.
11. Understanding the financial aspects of digital pathology: A dynamic customizable return on investment calculator for informed decision-making Ardon O, Asa SL, Lloyd MC, Lujan G, Parwani A, Santa-Rosario JC, Van Meter B, Samboy J, Pirain D, Blakely S, Hanna MG. Understanding the financial aspects of digital pathology: A dynamic customizable return on investment calculator for informed decision-making. J Pathol Inform. 2024 Apr 10;15:100376. doi: 10.1016/j.jpi.2024.100376. PMID: 38736870; PMCID: PMC11087961.
12. Diagnostic Technology: Trends of Use and Availability in a 10-Year Period (2011–2020) among Sixteen OECD Countries Martella M, Lenzi J, Gianino MM. Diagnostic Technology: Trends of Use and Availability in a 10-Year Period (2011-2020) among Sixteen OECD Countries. Healthcare (Basel). 2023 Jul 20;11(14):2078. doi: 10.3390/healthcare11142078. PMID: 37510518; PMCID: PMC10378781.
13. Costs, charges, and revenues for hospital diagnostic imaging procedures: differences by modality and hospital characteristics Sistrom CL, McKay NL. Costs, charges, and revenues for hospital diagnostic imaging procedures: differences by modality and hospital characteristics. J Am Coll Radiol. 2005 Jun;2(6):511-9. doi: 10.1016/j.jacr.2004.09.013. PMID: 17411868.
14. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop Curtis P. Langlotz, Bibb Allen, Bradley J. Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S. Cook, Adam E. Flanders, Matthew P. Lungren, David S. Mendelson, Jeffrey D. Rudie, Ge Wang, and Krishna Kandarpa Radiology 2019 291:3, 781-791