ACR®
AMERICAN COLLEGE OF
RADIOLOGY

AI-LAB Reference Architecture Framework

Public Comment

Editors:
Chris Treml, ACR
Brad Genereaux, NVIDIA

This document defines the reference architecture framework for AI-LAB. AI-LAB is aimed to be used by hospitals, clinics and imaging centers seeking to create, assess, experiment with, and leverage AI algorithms within their institution. It also provides supporting literature for infrastructure sizing, solution evolution, and other considerations.

The intended audience are those who are responsible or contribute to the hospital IT ecosystem; this includes hospital IT analysts and architects, program managers and leadership, healthcare software vendor architects, platform specialists, and IT system purchasing.

Contents

1. Introduction

A healthcare institution wishing to include artificial intelligence (AI) workflows in medical imaging must design and implement an architecture to support it. This architecture supports the workflows of inference, model validation, model training, and study annotation, and scales according to the needs of the institution. The infrastructure can grow into one with capabilities for continuous improvement, monitoring, evaluation, and federated learning regardless of immediate need.

In order to deliver these capabilities, AI Compute infrastructure is architected into the existing hospital data center infrastructure.

The main design goals of such an architecture must be:

- **Unified**: a single system infrastructure that is easy to manage across the enterprise
- **Scalable**: a system that can grow with the use cases and study volumes
- **Fault-tolerant**: a system where an individual failure does not impede healthcare delivery
- **Repeatable**: derived from a blueprint relevant to all types of healthcare institutions

Success criteria of this architecture demonstrates the following qualities:

- **Connected workflow:** Ability to configure imaging ecosystems to transfer medical images and metadata between AI architecture and PACS, RIS, VNA, EMR, and DICOM routers
- **Inference pipeline:** Ability to execute analysis pipelines to perform inference (e.g. segmentation of cardiac wall, and estimation of quantitative measures)
- **Shared AI and federated learning:** Ability to download models from trusted sources, augment them with local annotated data, and share them with others
- **Operational harmony:** Ability to articulate how the ecosystem is supported by hospital IT departments and is enterprise-ready

1.1. General Hospital Topology with AI Compute Infrastructure

The following diagram shows a typical hospital data center environment that includes existing EMRs, PACS, database, storage, and networking clusters, in addition to a new cluster for AI Compute.

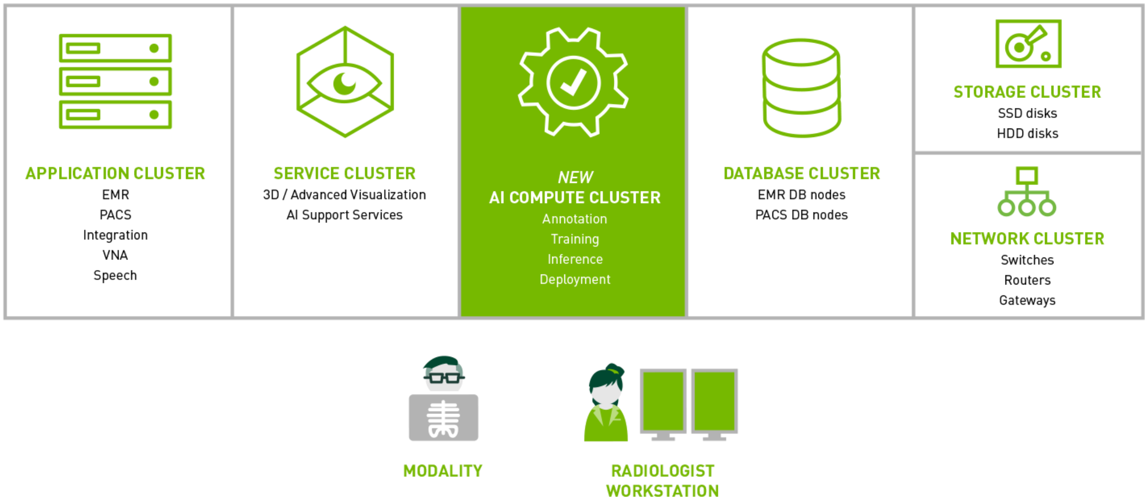

Figure 1: General hospital topology with AI Compute infrastructure.

AI Compute infrastructure is complementary to a hospital medical imaging ecosystem, without replacing any of its components. PACS, VNA, interface engines, advanced visualization services, and AI analysis platforms remain in place, with AI Compute augmenting the services already present.

1.2. Functions of AI Compute Infrastructure

The workflows that an AI Compute infrastructure will need to participate in are described in Chapter 2 - but the high-level functions include model and annotation transfer and storage, annotation creation and processing, training and fine-tuning models, validation and inference using models and imaging data. The AI Compute infrastructure may need to connect to existing systems including PACS, VNA, modalities, reporting systems, and EMRs. In addition, in order to participate in shared model creation and learning, global connectivity to model repositories will be necessary. The following diagram depicts the AI Compute infrastructure integrated with existing systems.

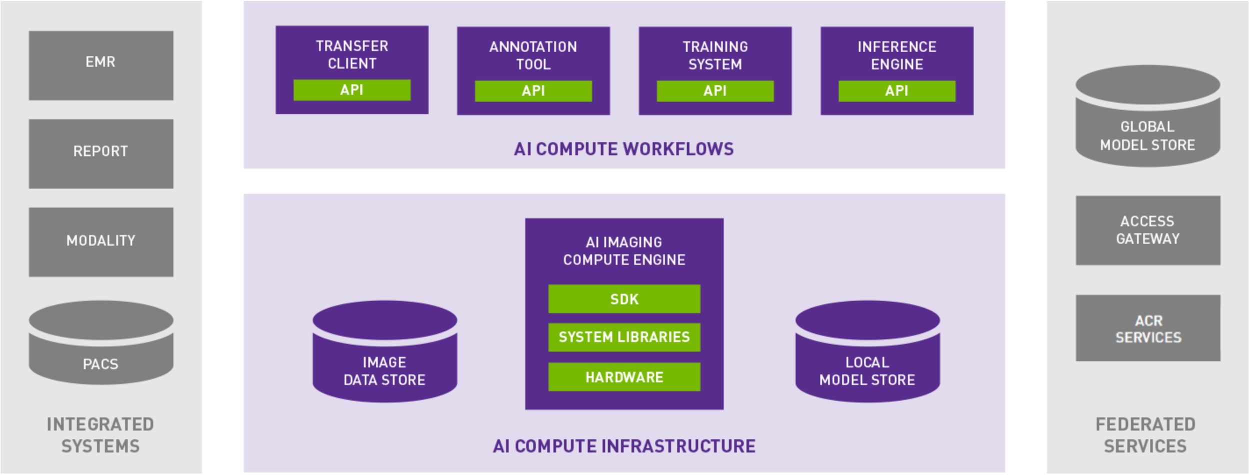

Figure 2: Key components of AI Compute infrastructure.

2. Know the AI Workflows

Before AI Compute infrastructure can be sized and scoped, it is important to understand what workflows will be performed at the healthcare institution and the estimated volumes that will be performed by these workflows. The following sections describe the key workflows that can be performed.

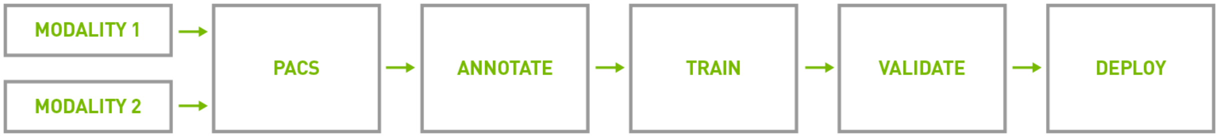

Figure 3: Simplified AI workflow chain.

2.1. Annotate

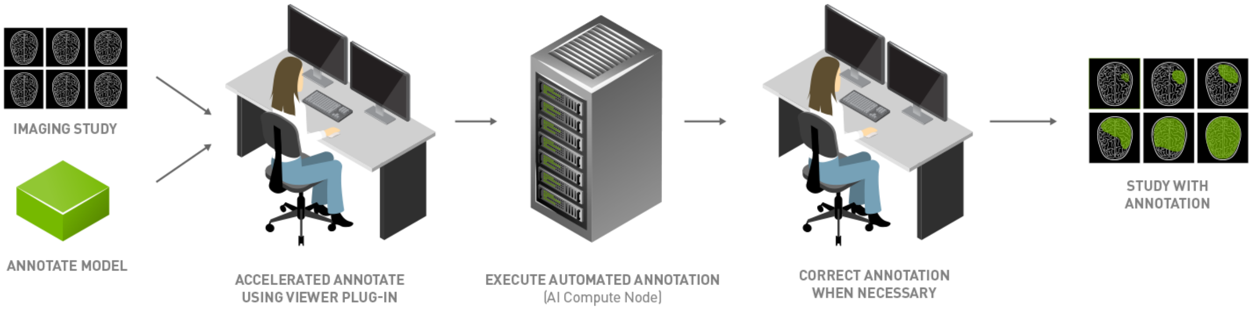
IMAGING STUDY
ANNOTATE MODEL
ACCELERATED ANNOTATE USING VIEWER PLUG-IN
EXECUTE AUTOMATED ANNOTATION (AI Compute Node)
CORRECT ANNOTATION WHEN NECESSARY
STUDY WITH ANNOTATION

Figure 4: Rapid creation of annotated datasets using annotation models.

The following items are needed for annotating images in an accelerated fashion:

- AI Compute infrastructure to run AI-assisted annotation pipelines
- A set of imaging studies appropriate for annotation
    - To generate these, clinicians require access to PACS or an imaging archive
- Appropriate annotation models
- An annotation toolkit

2.2. Train

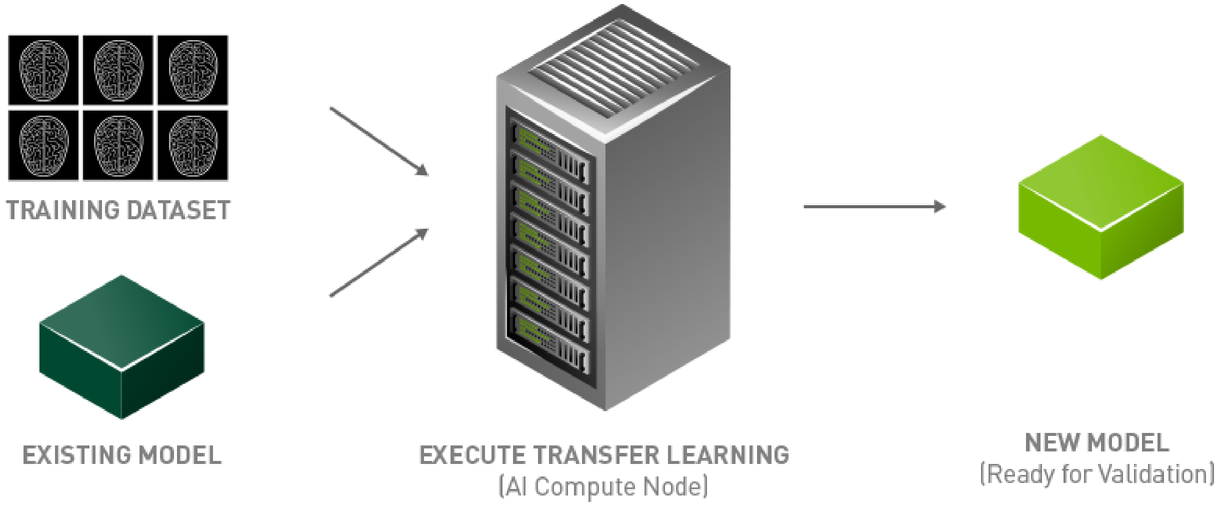

Figure 5: Transfer learning using a local annotated dataset to augment a model downloaded from a global repository.

The following items are needed for augmenting algorithms with locally annotated data:

- AI Compute infrastructure to run inference pipelines
- A set of annotated imaging studies suitable for training a model
    - To generate these, clinicians require access to an annotation creation tool
- An existing model to be augmented
- A transfer learning toolkit to execute model transfer learning

2.3. Validate

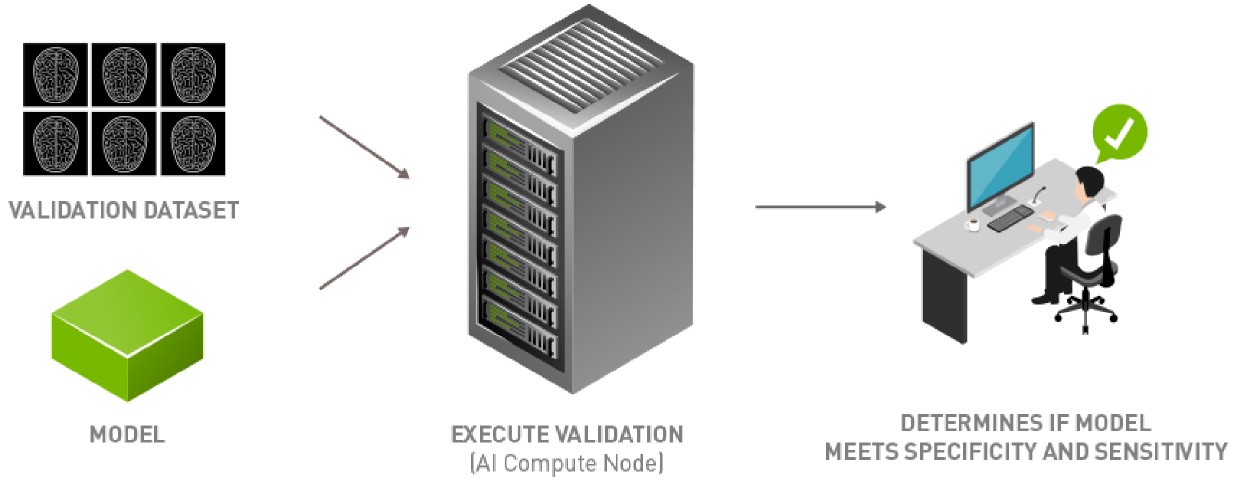

Figure 6: Model download from a global repository and validating against a local source of imaging studies.

The following items are needed to validate an algorithm:

- AI Compute infrastructure to run inference pipelines
- A set of annotated imaging studies suitable for validating a model
    - To generate these, clinicians require access to an annotation creation tool

- A model to validate
- A transfer learning toolkit to execute model validation

2.4. Infer

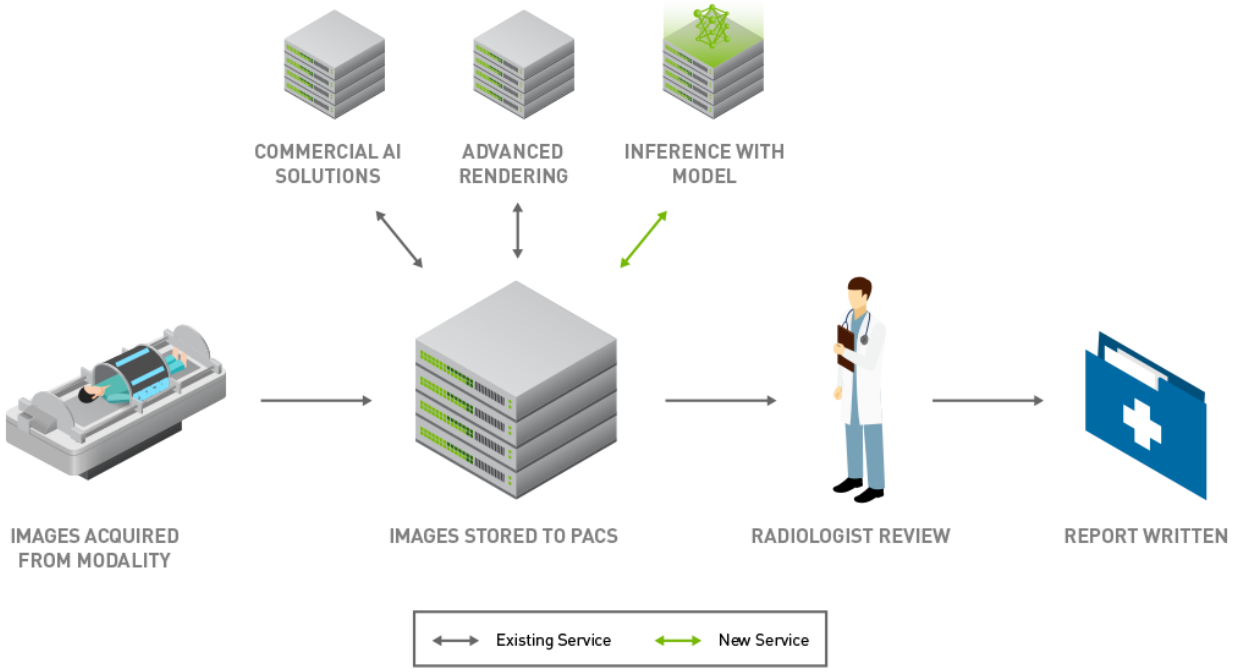

Figure 7: Imaging study workflow from image acquisition to report generation.

The following items are needed to set up inference of a model on imaging studies:

- AI Compute infrastructure to run inference pipelines
- PACS configuration to add a routing destination for imaging studies
- A DICOM adapter to receive DICOM instances and send results
- Pre-trained model to run DICOM images through
- An inference server to execute the models
- PACS configuration to accept results from the DICOM adapter

3. Sizing AI Compute Infrastructure

AI workloads - in particular, model training, are computationally heavier than traditional applications such as EMRs and PACS, and the databases supporting those systems. As such, it is important to understand the processing power needed, and how to scope, scale, and grow an environment to support these loads.

3.1. Determine Approximate Size

Using the workloads defined in Chapter 2, the size and profile of the AI Compute infrastructure can be determined and generally categorized as follows:

- **PILOT:** A clinician or small research team that wants to start small, and does not have many resources to start
- **SMALL:** A small hospital or research group that intends to focus on validation and inference but not a lot of model development
- **MEDIUM:** A medium hospital or research group that employs or plans to employ a data scientist or small team
- **LARGE:** A large hospital or research group that has a data science team or is connected to an academic institution

Some healthcare institutions may have more specific needs, based on their specific workloads that they are running. The following chart provides some more specific tuning to the size of compute needed.

Workload	**Measure**	**Average Size Needed**
Annotate	No annotation being performed	N/A
	Annotating less than 10 exams an hour	Small
	Annotating more than 10 exams an hour	Medium
Train	No training being performed	N/A
	Fine-tuning existing models	Medium
	Training new models	Medium-Large
	Training multiple models simultaneously	Large
Validate	No models being validated	N/A

	Validating less than 30 exams an hour	Small
	Validating more than 30 exams an hour	Medium
Infer	No inference being performed	N/A
	Inferencing less than 30 exams an hour	Small
	Inferencing more than 30 exams an hour	Medium

The general size is only an approximation; refer to Chapter 4 for additional criteria considerations that may influence sizing.

3.2. Sizing Typical Configurations

Size	TFlops Range	Sample Footprint	Examples
Pilot	<15	Desktop or laptop	Off the shelf desktop computer can be used for single-threaded inference tasks with a limited capacity
		Cloud Instance	Microsoft Azure Dsv3-series sizes are suitable for light inference tasks and model validation where completion time isn't a factor
		Desktop or laptop with GPU	A laptop powered with an NVIDIA Quadro GPU is enough to get started with basic AI activities
Small	15-480	Data science workstation	A workstation powered by NVIDIA Quadro RTX 6000 is suitable for some model fine-tuning and associated workflows
		Cloud Instance with GPU	Microsoft Azure NDv1-series VMs are powered by NVIDIA Tesla P40 GPUs suitable for remote model learning and fine-tuning workflows
		Server powered by GPU	A server powered by NVIDIA T4 GPU or an NVIDIA DGX Station can be used to minimize latency and improve performance
Medium	480-2000	Server powered by GPU and optimized for data science	A server powered by NVIDIA Tesla V100 (for example, a DGX-1) would be suitable for training new algorithms
Large	2000 +	Server stack powered by GPU and optimized for data science	A server stack or multiple stacks powered by NVIDIA Tesla V100 (for example, a DGX-2)

Aside from looking at the processing footprint, there are other factors to consider. For example:

- RAM: AI libraries may dictate a minimum amount of RAM needed; e.g. 16GB or higher.
- Disk space: The amount of disk space required depends upon several factors;
    - Number, type, and size of annotated studies being stored
    - Number and versions of models and transforms being stored
    - Temporary space for study processing and queuing (dependent on number, type, and size of studies)
- Network: Dependent on data volumes; 1 Gbps or higher

3.3. Understand Service Experience

The following table defines several typical profiles and considerations.

Workload Size	**Experience**
Pilot	Systems not managed within the data center may not have appropriate business continuity or uptime for production. Expect that compute-heavy processes may take a much longer time to complete.
Small	Suitable for single runs of compute-intensive workflows but may slow down if load increases.
Medium	Suitable for most workloads with a mix of training, annotation, validation, and inference.
Large	Suitable for the largest of workloads and should be linearly scalable with additional systems when simultaneous loads become too great.

3.4. Sample Workload Scenario

For this example, Andrew (a fictitious radiologist with an IT background) is working for General Hospital (a fictitious hospital). Andrew is looking to start an AI program within their institution. He wishes to annotate 100 heart wall studies from chest CT exams, in order to fine-tune a general model shared by another institution. Once he has completed the fine-tuning, he will validate the tuned model with 20 studies from his PACS. Once satisfied, he will then conduct inference on approximately 30 studies per day.

Given these numbers, the following are some performance expectations:

Workload	**Metric**	**Sizing Configuration**	
		Pilot	**Small**

Annotate	Annotate 100 chest CT studies (serially)	Hours	<1 Hour
Train	Fine-tune 1 model with 100 annotated chest CT studies	Hours	<1 Hour
Validate	Validate model with 20 studies (serially)	<1 Hour	<1 Minute
Infer	Infer 30 studies (serially)	<1 Hour	<1 Minute

Timing only reflects processing time, and not the human time component to execute these tasks.

4. Design Factors when Building AI Compute Infrastructure

4.1. Factors that Influence Sizing

Given the general guidelines discussed in chapter 3, there are several additional factors that influence how to size the solution. These include:

Factor	Low-end Example	High-end Example
Image size	A CT image	A mammography image
Number of images per study	X-Ray	CT or MRI
Whether 2D or 3D analysis is being conducted	Ultrasound	CT or MRI
Average number of studies an institution generates	25-100 studies per day	300+ studies per day
Type of algorithm	Characterization, secondary capture	Segmentation, markup
Number of data scientists training algorithms at one time	Clinician experimenting with training	Team of 10 full time data scientists

4.2. Data Security and Privacy

Patient sensitive information must always be protected, both while in transit and at rest. Ensure that there are adequate data protections throughout the entire lifecycle.

Building out AI Compute infrastructure on-site ensures that sensitive patient data remains within the local hospital network. If using a cloud provider, one must additionally consider how to transmit and receive data securely and recognize that a Business Associate Agreement (BAA) may need to be in place with the cloud provider, and that the provider is certified to store health data using HIPAA policies.

4.3. Model Management and Hospital System Integration

Initially, hospitals may choose to utilize local disk storage for managing models, transforms, annotated data sets, and inference results. For example, having a local file directory for annotated studies for a particular model may be acceptable for informal use for one user, but this is not necessarily scalable when there are multiple users sharing the same infrastructure. In this case, enterprise management of this content is important.

In addition, when building out inference pipelines, and connected into a medical imaging repository, like PACS, considerations must be made on how to get data in and out of those systems. For example, in a single user experimentation use case, it may be fine to export DICOM studies from a PACS to a local disk drive; but for workflows in a pipeline, DICOM routing through the use of classic DIMSE (C-MOVE, C-STORE) or DICOMweb (WADO-RS) protocols will be much more important.

Pipelines may also require the use of data transformation layers - for example, converting a DICOM study consisting of a series of CT images into a NIFTI representation is a typical necessity.

4.4. Data Beyond Images

There may be information in the broader ecosystem that needs to be considered as part of the AI analysis. For example, having a disease problem list, understanding past surgeries (e.g., amputations, organ donations, etc.), and impacts of co-morbidities or pharmaceuticals may influence the outcomes of AI operations.

However, providing access to the broader health data repositories (e.g., the EMR) has its own set of complications, such as:

- Legality
- Privacy
- Access
- Version-specific availability of APIs and update lag-time for sites deploying new releases of software

4.5. Disaster Recovery, Monitoring and Business Continuity

Once a system is productized, it must also be reliable - meaning, if a node goes down (power outage, hardware or software failure, etc.), hospital business can proceed unencumbered.

The ecosystem should be able to be monitored for uptime, performance issues, and impending problems (such as running low on disk space or maximizing CPU/GPU usage).

4.6. Business Intelligence

It is important to measure the impact of leveraging AI in the ecosystem, so that stakeholders understand the value and further opportunities. E.g., knowing things such as:

- How many studies are being processed?
- How long does it take each study to process?
- How much time is this saving the institution?

- Where is the next biggest opportunity to apply AI in the hospital?

4.7. Data Lifecycle Management

Understand the requirements from a health record perspective on whether to keep AI-generated data. Local laws may require if a clinician reviewed or used the data as part of clinical practice, that data must be retained as part of the imaging record, which may impact liability, storage commitments, and more.

It may also be necessary to retain the specific versions of models used as part of inference, in order to re-execute the inference in the future for liability purposes.

Keeping the annotated data sets used to train a model (or augment further training of a model) is important. Annotated data sets remain valuable assets as they take time to create. There may be regulatory requirements to retain data sets and tag them based on which models they were applied against.

4.8. Ongoing Quality Management

When performing inference, it is important that models are monitored over time to ensure that their specificity and sensitivity remain within acceptable parameters - similar to peer review. As clinics and techniques evolve over time, there may be changes that impact the inference process, including:

- Adding new modalities
- Updating the firmware of a modality
- Re-calibrating a modality
- Accepting studies from outside the institution from other modalities

In addition, techniques such as continuous learning (improving the models based on production reporting output), and federated learning (leveraging the collected learning of many sites) will also require ongoing quality monitoring and management.

5. Defining an Evolution Strategy

A hospital adopting AI into workflows takes time, learning, and investment, as discussed in the introduction. To get there, hospitals embark upon an iterative journey, transforming workflows in proof of concept settings, broadening into pilot projects, and ultimately moving models into production use. At each stage in this journey, there are specific goals to reach.

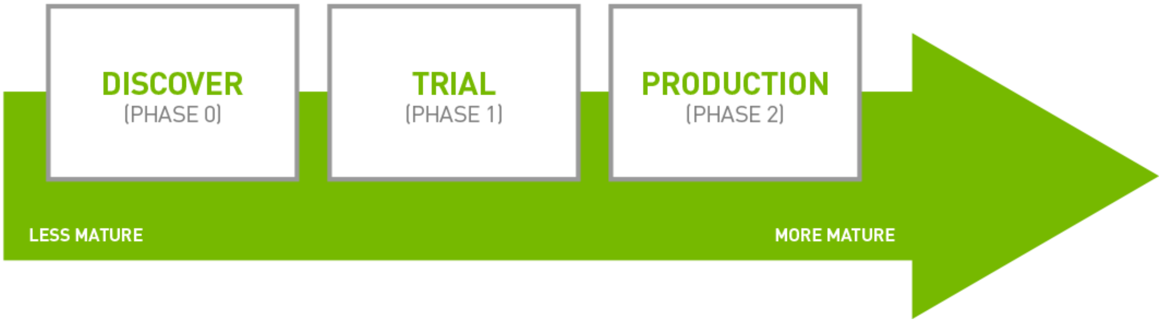

Figure 8: The three steps in AI Compute evolution strategy.

Beyond phase 2 moves beyond the data science and introduces AI directly into clinical workflow. There are many things to additionally consider, from liability to medical device to communication to acceptance, that goes beyond the scope of these first three phases.

5.1. Phase 0: Discover

GOALS:

- Identify challenges that the hospital faces today that AI can address
- Establish governance with a cross-disciplinary team to champion AI in the institution
- Learn about the impact of AI on workflow and infrastructure
- Gather details about the ecosystem readiness

Cost +

Resources + +

In this phase, the focus is on preparatory work. A hospital should take stock of the challenges and opportunities AI can address. For example:

- Where are radiologists spending the most amount of time?
- What conditions and diseases are seen at the hospital where current workflow is less than optimal?
- What datasets exist that can be used to get started?

The next step is to identify a cross-disciplinary team to understand the vision for building AI solutions within the hospital. This group typically consists of a clinical champion (e.g., a physician or radiologist), an informaticist champion (e.g., a PACS administrator or a modality technician), and an IT architect champion (e.g., a systems analyst) - and if possible, a data scientist or statistician.

The team needs to understand the principles of AI and how to leverage AI in their environment. In addition to this document, there are many other resources available. Educational events like RSNA[1], C-MIMI[2], and GTC[3] allow attendees to learn about libraries and possible architectures that work within a hospital. Hands-on workshops allow attendees to annotate studies and see the results of inference. Many courses and related content are available online, including from the ACR, NVIDIA DLI, Coursera, fast.ai, and others.

5.2. Phase 1: Trial

GOALS:

- Identify specific AI workflow to trial at the hospital
- Build out small-footprint infrastructure or secure cloud space
- Connect data sources from a test environment (PACS, VNA, EMR)
- Run models against test data and analyze results
- Perform transfer learning to improve results

Cost +

Resources + + +

In this phase, the focus is to set up an environment to trial execute AI workflows in a low-risk, research-like way, in order to demonstrate the value of building a production-class infrastructure in future.

The investment in the environment can be small at first, depending on the types of models being run. The intent is to create a test and proof-of-concept environment, and not on real-time performant analysis. Models could also be run in a cloud instance connected to local data sources - but care must be taken around protecting any protected health data.

When choosing an AI workflow to trial at the hospital, it may be worthwhile to review what existing reference models are available.

For this phase to be successful, validation of algorithms with an acceptable level of specificity and sensitivity must be shown. There should be significant evidence of value to build a production-type environment and justify the hard costs of investing in the architecture (rather

[1] See https://www.rsna.org/.
[2] See https://siim.org/page/2019cmimi.
[3] See https://www.nvidia.com/en-eu/gtc/.

than soft costs of people investment and minimal cloud costs for trial flows). It should be possible to present a compelling business case to the hospital's board of directors or other stakeholders, showing the value of a platform built in a scalable, reusable, future-proof way.

5.3. Phase 2: Research Production

GOALS:

- Develop production architecture appropriate for workflow and institution size
- Plan operational support and service-level agreements for the infrastructure
- Educate the clinical and support staff of the new functionality available to them, and cover appropriate uses
- Develop a roadmap for which models to introduce next, and include transfer learning as part of the process

Cost + +

Resources + +

In this phase, once the value of a proof-of-concept is demonstrated, a hospital system may be ready to bring AI workflows into production and scale out the system infrastructure. As with any infrastructure initiative, ensuring a strong governance model is key; this is a joint effort between clinicians, IT, and data scientists.

A hospital may choose to build out a system on-premise to resolve the question of scale. If image processing is done in the cloud, there is a requirement that the imaging studies be pushed to the cloud prior to processing, which incurs latency (as studies are transmitted to and from the cloud), reduced Internet bandwidth for other traffic, added operational cost (to use cloud storage, and possible risk of downtime (if Internet connectivity is lost).

Considering that this environment is to be used in production workflows, it becomes crucial that, just like any other health IT system, all aspects of uptime are considered. For example, setting a target of 99.9% uptime gives nearly 9 hours of downtime per year. To achieve this level of service, the following should be considered:

- Redundancy (an entire node can fail without impacting workflow) and fault tolerance (individual components like a power supply can fail without impacting workflow)
- Monitoring (for both service availability and potential problems like low disk space or high temperature)
- Hardware and operating system patches and updates

Bringing an AI Compute infrastructure into production absolutely requires institutional communication, training, and setting the right level of expectations for clinicians on how to use the models as part of their clinical workflow.

Developing a roadmap of what additional models to bring online is important - as new models become available, and technologies and systems evolve. The infrastructure used for production, if scaled and planned appropriately, should be able to meet the needs of data scientists to continue in their validation and training work without impacting clinical practice whatsoever.

For this phase to be successful, it must be demonstrated that end-to-end processing takes place, satisfying the high levels of specificity and sensitivity, while at the same time satisfying performance requirements (keeping up with the loads) and the needs of IT administrators for system resilience.

6. Attribution

Images and figures courtesy NVIDIA Corporation.