Lean Study Host: Towards an Automated Pipeline for Multi-Center Study Hosting

Lukas Heine
University Medicine Essen
lukas.heine@uk-essen.de

Fabian Hörst
University Medicine Essen
fabian.hoerst@uk-essen.de

Enrico Nasca
University Medicine Essen
enrico.nasca@uk-essen.de

Jan Egger
University Medicine Essen
jan.egger@uk-essen.de

Jens T. Siveke
University Medicine Essen &
German Cancer Consortium (DKTK)
jens.siveke@uk-essen.de

Moon Kim
University Medicine Essen
moon.kim@uk-essen.de

Jens Kleesiek
University Medicine Essen &
TU Dortmund University
jens.kleesiek@uk-essen.de

Fin H. Bahnsen
University Medicine Essen
fin.bahnsen@uk-essen.de

Abstract

Medical studies are an essential part of advancing research. A uniform, flexible software infrastructure that allows for straightforward data management stands at the core of studies that involve multiple sites. Such a solution must accommodate the specific technical needs of clinical practitioners and researchers, such as uploading, viewing, downloading, annotating, and sharing image material in various forms. The current tool landscape needs a solution that bridges the gap between intuitive data governance and usability without introducing undesired technical and legal overhead. We present "Lean Study Host" (LSH), a novel, open-source approach to clinical study data management that caters to clinicians, technical staff, and data protection officers. It seeks to reduce technical, administrative, and legal overhead to allow studies to focus more efforts on research. It combines a cloud-native, microservice-based architecture, deidentification, and on-premises hosting to keep data sovereignty within the local institution.

Keywords: Study management, Healthcare infrastructure, CI/CD, FHIR, Data management platform

1. Introduction

Evidence-based medicine is advanced through research and studies. Images acquired from different modalities and multiple hospitals are a major part of them. Image-based studies have long been the standard in various medical fields (Patel et al., 2011, Lewin et al., 2007, Fayad and Fuster, 2000), as they serve diagnostic purposes and allow for therapy monitoring.

Computer Tomography (CT) or Magnetic Resonance Imaging (MRI) recordings, are valuable sources of information throughout the entire treatment process. Advances in the field of machine learning gave rise to increasingly complex algorithms (Sun et al., 2020, Hörst et al., 2023, Fleagle et al., 1989) that rival the performance of human specialists, if given sufficient data. This data is usually collected through a collaboration of multiple sites, which can lead to conflicting interests. Efforts to encourage data sharing may not be successful in practice (Watson et al., 2022), which may lead to the unwarranted repetition of studies.

The current software landscape lacks open-source software that addresses problems that commonly arise in such a scenario in a satisfactory manner. We propose "Lean Study Host" to tackle these challenges. The overarching goal of LSH is to help medical researchers uncover widespread, emergent patterns affecting public health and how they can be addressed optimally. To investigate this more dynamically, an easily deployable, customizable, and extensible solution for conducting medical, image-based studies is of utmost importance. In this context, LSH serves as a universally deployable platform that can be used as is or extended to fit a specific use case of a study. At the same time, it provides means of user group management, monitoring changes made to study data, and auditability to be fully compliant with legal requirements. Our approach leverages standardized, tried, and tested open-source software and the most prevalent medical standards and protocols to build a robust, reliable system.

In this submission, we compile the requirements for study software. These requirements are further influenced by legal matters and aspects such as data protection and data governance. Our contribution fulfills these requirements holistically. From an information technology perspective on the problem, LSH relies on

URI: https://hdl.handle.net/10125/106775
978-0-9981331-7-1

modern technologies such as containerization, through which we gain a high degree of scalability in the cloud-native approach and a high degree of transparency through the Git-driven deployment architecture.

The paper is structured as follows: In section 2, we describe the research context and outline works that influence our work, followed by the state-of-the-art 3. In section 4, the building blocks of LSH are described, while section 5 describes the actual use cases concerning the conduction of medical studies with LSH. Finally, we discuss compliance matters in section 6 as well as its limitations in section 7 and conclude the paper in section 8.

2. Research Context & Considerations

From an organizational standpoint, data platforms at the core of such undertakings can be defined as "multi-stakeholder arrangements for the organization of data storage, processing and sharing." (Gubser et al., 2023, p. 1). Data governance stands at the core of these agreements "to reconcile conflicting interests in data (use), which diverge amongst the different stakeholders involved in terms of its value and risks." (Grafenstein, 2022, p. 5). The different data governance entities can be organized in three layers: the normative layer, the organizational layer and the technological layer (Grafenstein, 2022). LSH addresses data governance in the technological layer. Consequently, it relies on decisions made by stakeholders on the other layers. Using sensitive healthcare data entails a tradeoff between research interests and patient-related data privacy: On one hand, the information contained in images is crucial for the development of new treatments and can serve important purposes for data modeling. On the other hand, it is very sensitive data as it describes the very nature of the patient. This sensitivity has long been recognized as any kind of image-based information, and its metadata are considered Patient Health Information (PHI), which is protected by the European General Data Protection Regulation as well as the U.S. Health Insurance Portability and Accountability Act. Despite numerous works that highlight the legal, organizational and technical differences in governance of data platforms, there is a distinct lack of technical solutions that are flexible enough to meet changing demands. If multiple sites cooperate, a multilateral agreement is required to document how patient information is correctly handled.

A primary source of concern in the domain of image-based studies stems from metadata contained in so-called tags defined by the Digital Imaging and Communications in Medicine (DICOM) standard. These tags contain a multitude of information, ranging from device serial numbers to a patient's birth data. On one extreme of the tradeoff, any metadata attached to the image would be redacted. However, this practice may contradict the requirements of a study since certain background information may be required to draw conclusions about a pathology. A patient's sex, age, and ethnicity may be highly influential in determining the treatment outcome but can also be used to help identify a patient. Consequently, each study must decide how to treat every individual tag to preserve the information necessary for the research objective. In contrast, any information that violates legal constraints must be redacted.

This becomes even more relevant when the data collected in a study is used to train machine learning algorithms or in other downstream applications. Besides privacy concerns, other aspects need to be taken into account as well. A concept that is often mentioned in this domain is the FAIRness of data (Wilkinson et al., 2016), which means that it is findable, accessible, interoperable and reusable. These criteria play an important role, especially in downstream tasks that involve machine learning or data science applications. Interoperability standards such as Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) (Bender and Sartipi, 2013) can help address this.

The European Commission recognized several of these challenges and addressed them in a legislative proposal titled "The European Health Data Space" which was subsequently analyzed in a corresponding research paper (Marcus et al., 2022). Among other findings, the authors conclude that "Privacy for secondary use can be achieved through anonymisation. Decentralized data pools at national level might further contribute to privacy" (Marcus et al., 2022, p. 8). Further, they recommend the unionwide use of FHIR to aid interoperability. FHIR can help with more than just interoperability: an API connected to a dedicated server also serves documentary purposes. A real-time catalog of stored resources is available through a server's REST API. These implementation efforts are driven by recommendations on data cataloguing (Grafenstein, 2022) and the need to extend the availability of self-hostable, FHIR-supporting data platforms to a wider audience (Davidson et al., 2023).

To the best of our knowledge, a software that addresses all of these concerns is not yet available in clinical use. As such, we propose LSH to align the knowledge of existing work with new legislative guidelines and insights gathered from research.

Table 1. Requirements, solutions and comparison with other tools

Requirement \ Software	LSH	REDCap	Nora	XNAT	JIP
Open-source	MIT License	✗	✗	✓	✓
DICOM viewer	OHIF-based	✗	✓	✓	✓
Deidentification	DICOM standard-guided, based on deid	✗	✗	✓	✗
WSI support	Conversion is OrthancWSIDicomizer based, visualization uses OpenLayers	✗	✗	✗	✗
Full FHIR support	Backbone is formed by HAPI FHIR	✗	✗	✗	✗

3. State of the Art

Other endeavors that precede this work influenced the design choices that were made during the development of LSH. One such work was presented by Shaban-Nejad et al., 2017 on population health information systems. During their research, the authors identified several reoccurring challenges. One of them is that "different data sources are heterogeneous with discrepancies that present challenges for their collection, integration, and processing [...]" (Shaban-Nejad et al., 2017, p. 45). To address this in our work, data harmonization can be configured on a per-rollout basis to avoid inhomogeneities as best as possible. As a consequence, pipelines can ensure that different forms of data get converted to formats that are deemed usable. All stored data adheres to predefined standards such as DICOM or HL7 FHIR. These measures enable data to be easily shared across instances or migrated to existing infrastructure. Contrary to their work, LSH does not focus on distributed learning systems but on the entire study conduction process as a whole.

Another approach to sharing data and applications responsibly was presented by Choudhury et al., 2020. The authors introduce the so-called *Personal Health Train* in their work, which serves as a framework for federated machine learning. However, they do not address the same, clinical research setting, but give recommendations on the handling of data (such as the FAIR principles). Another software that has found widespread use for the support of clinical studies is *Research Electronic Data Capture* (REDCap) (Harris et al., 2009). While there is some functional overlap with LSH's features, several differences will be outlined. First and foremost, our work puts explicit emphasis on image-based studies. Our implementation does not rely on forms and questionnaires that are manually filled to acquire data. Instead, we rely on data already present in common file formats. Moreover, our system is fully containerized and is more easily deployable. In contrast to tools like REDCap, our solution assumes no prerequisites apart from a container engine. Containerization significantly reduces the time and effort that needs to be put into the setup, among other benefits that will be covered in more detail in the next section. Furthermore, our implementation includes a fully functional, dedicated PACS that offers researchers an intuitive way of interacting with image data. Another medical research tool is the *Joint Imaging Platform* (Scherer et al., 2020) (JIP). In their work, the authors describe another approach to managing imaging data. In contrast to other solutions, JIP recognized the need for systems like a PACS and integrated it into their ecosystem. In comparison, LSH integrates a FHIR server to document the study's resources as well and emphasizes extensibility and interchangeability.

Studies can differ significantly in their needs, such as individual apps, custom processing, harmonization pipelines, and viewers. LSH embraces the heterogeneous nature of image-based studies and decides against a static solution: LSH strives to be minimalist by only supplying the tools that are required but increase the accessibility of exposed services to be able to deploy custom extensions if needed rapidly. In consequence, deployments are easy to set up and maintain. There are other data platform initiatives such as *XNAT central* (Herrick et al., 2016), *IDA* (Crawford et al., 2016) and *Vivli* (Bierer et al., 2016) that serve the purpose of making health data accessible to the public. LSH serves the purpose of representing an intermediate stage where data can be shared before being made available to large data repositories: Data collected in the clinical routine may not be immediately fit to be shared as it may not be deidentified, cleaned or sorted. After data has been collected in LSH, processed and cleared for public use, it can then be made available to supra-institutional data repositories. Other approaches that rely on distributed ledger technology (DLT) exist in this landscape as well. While employing blockchain technology in this setting does merit certain benefits (Beyene et al., 2022), they are fundamentally different in their architecture and use case and are thus out of the scope of this work. In summary, the implementation of standards that are soon to be made mandatory in the

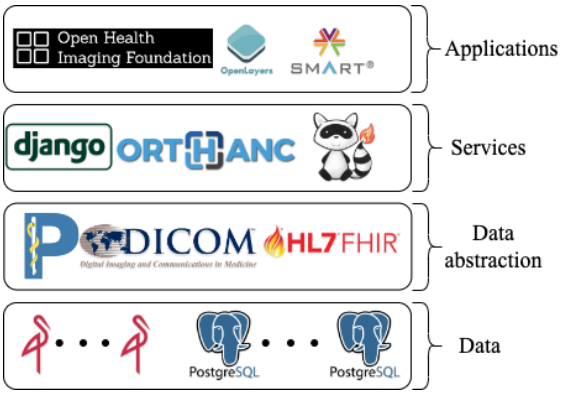

Figure 1. Lean Study host organizes individual components in layers which form a building block that is easy to manage

European Union, the integration of well established clinical software as well as design decisions set it apart from other open-source software. Requirements for such a platform were derived from research, clinical practice and other implementations in this domain. An overview of requirements as well as how LSH addresses them is given in Table 1. Moreover, we compare it with other established tools (REDCap Harris et al., 2009, Nora Anastasopoulos et al., 2017, XNAT (Herrick et al., 2016), JIP (Scherer et al., 2020)). This makes LSH a cutting-edge solution that addresses the challenges of medical studies holistically and novelly.

4. Building Blocks of the Implementation

LSH consists of different building blocks that form the tool stack. An overview of the components is given in Figure 1. The following chapter explains how certain core concepts, as well as the mentioned building blocks, comprise the LSH ecosystem.

4.1. Containerization

Containerization is the process of packaging an application along with its dependencies, configuration files, and other necessary components into a lightweight and portable environment called a container. Containers provide an isolated and consistent environment for running applications, making it easier to deploy and manage them across different platforms by reducing the overhead of managing and installing dependencies or debugging system-specific errors. Numerous key benefits give rise to the use of containers in this case. One such benefit is portability: containers can run on any platform that supports containerization, making it easier to move applications across different environments without the need for significant changes. This is of utmost importance as different hardware and operating systems are prevalent in the research landscape. Another benefit is scalability: containers can be scaled up or down quickly and easily to meet changing demands, making them ideal for applications with varying workloads. Moreover, the efficiency of the system can benefit as containers are lightweight and require fewer system resources than traditional virtual machines, enabling greater efficiency in resource utilization. Last but not least, security can be improved as well. containers are designed to be secure by default, with built-in features such as isolation, sandboxing, and resource limitations to prevent unauthorized access and attacks.

4.2. Access and User Management

It is necessary to restrict certain privileges, such as viewing or modifying existing data to specific users or user groups. Further, rigorous documentation regarding these operations must be in place to monitor changes and access data tightly. *Teleport* greatly simplifies this by utilizing an Identity-Aware Access Proxy every connection passes through. Accordingly, clear and central user management is crucial. LSH relies on the use of *KeyCloak*, a commercially maintained open-source authentication provider. Finally, KeyCloak enables a uniform single sign-on mechanism for the different tools that can be composed in LSH.

Both tools fulfill different tasks, despite serving the same overarching goal: On the one hand, Teleport secures infrastructure by controlling and monitoring access to servers and other resources, while on the other hand, KeyCloak secures the individual applications running on said infrastructure by managing user authentication, access control, and identity federation. Therefore, users can be granted access to services and infrastructure based on their assigned roles. Moreover, these tools secure the host infrastructure with two-factor authentication (2FA). This adds an additional layer of security to protect sensitive data and infrastructure. This also helps with auditability as all accesses to monitored infrastructure can be logged and reviewed.

4.3. Picture Archiving and Communication System

A Picture Archiving and Communication System (PACS) is a medical imaging technology used for storing, retrieving, managing, and distributing medical images like CT and MRI scans as well as other DICOM standard based image material. It enables users to access and view patient images and information from any location, thus aiding diagnosis, treatment planning, and patient care. A PACS integrates with other healthcare

information systems and Electronic Health Records (EHRs) (as is the case with the LSH ecosystem) to provide a comprehensive view of a patient's medical history. Inside the LSH ecosystem, a PACS is a central access point for any DICOM data collected during the study. It can be configured to receive, filter, and send studies via DICOMweb to seamlessly integrate with existing PACS solutions. This means that most acquisition modalities in clinical use can be connected to the PACS instance running in the LSH ecosystem to allow data transfer directly from the device or another PACS instance to the study platform.

Another vital component accessible via the PACS is an *Open Health Imaging Foundation* (OHIF) viewer instance (Ziegler et al., 2020) that can display DICOM images. Moreover, clinicians can annotate and comment on selected images, making it easier to share and document image- or patient-specific information. In LSH, the recommended implementation of a PACS is formed by *Orthanc* (Jodogne, 2018).

4.4. Storage and Backups

The data storage is formed by *MinIO* S3 buckets that are deployed alongside LSH. For each running bucket instance, a backup bucket inside a separate container is deployed that mirrors the contents of the production container. In case of a malfunction or unexpected service outage, data can be restored from backup copies. More than one redundant backup can be stored, depending on the expected volatility of the employed hardware. Access to MinIO buckets can be granted to developers to access raw data for, e.g., machine learning applications. Various events can cause the loss of data or the corruption of the storage medium, the cause of these events can range from hardware faults to human error or other sporadically occurring events.

4.5. Data Management

Large cohort sizes, as well as multiple data sources, require scalable solutions to track resources associated with a study participant as well as relevant metadata. LSH's central information management is built around FHIR. FHIR is an open standard for documenting and exchanging healthcare information electronically. It was developed by the health IT standards organization, Health Level Seven International (HL7), and is the modern, established standard for healthcare data exchange. It is built on ubiquitous web technologies such as RESTful APIs, JSON, and XML, making implementing, extending, and integrating with existing systems easier. The FHIR standard promotes interoperability, simplifies implementation, and improves patient care by providing a consistent framework for exchanging clinical and administrative data between healthcare systems. The RESTful API allows a straightforward entry point for all microservices requiring access to study information. In addition to that, FHIR serves documentary purposes for all collected data. For instance, the patient count, number of available images, and other resources can be retrieved quickly through simplistic HTTPS requests. In contrast to DICOM files, FHIR documents a subset of its metadata and stores it in its corresponding resource. Furthermore, FHIR is able to represent semantic structures such as imaging instances grouped in series that belong to an imaging study. This allows users to quickly understand relationships between different types of data.

Moreover, the FHIR server also serves important functions for downstream applications, especially for machine learning and data science-based ones. Applications can make use of *SMARTonFHIR* (Mandel et al., 2016) to get secure access to EHRs in the study. Requests made to the FHIR server aid in finding, filtering, and accessing the information contained in its stored resources. The results of requests made to the same server are reproducible, simplifying the interaction with data in a multi-centric setting. It thus forms an additional abstraction layer that allows users to work with the collected data more efficiently. Thus, the FHIR server functions as the *brain* of the data layer, as all resources connected to the study are documented in a central place. One implementation of a suitable server is HAPI FHIR. FHIR resource creation is triggered by the storages and orchestrated by a *Django* server component. Once a file is uploaded via the web UI or received from a connected DICOM modality, it is redirected to its designated storage location inside the MinIO bucket. Depending on the type of file, one of two workflows can be triggered: DICOM files are sent to the PACS and the corresponding FHIR resource on the server is updated whenever a new series instance is uploaded.

Non-DICOM files are uploaded into their designated bucket. Whenever a create, read, update, or delete (CRUD) operation is registered inside the bucket, the corresponding resource on the FHIR server is updated. There are two ways these notifications can be issued to the Django server. Webhooks are better suited for smaller, simple studies. Whenever a monitored event happens, a URL on the Django server is called that handles the resource CRUD overhead. Secondly, a message broker like RabbitMQ can be used in large-scale studies that need to support a larger, more complex ecosystem of applications.

4.6. Deidentification

LSH defines a default pipeline to deidentify DICOM files. Contrary to the term "anonymization", "deidentification" means the best effort towards anonymization. At the time of writing, the normative Attribute Confidentiality Profiles defined in DICOM PS3.15 2023b - Security and System Management Profiles Committee, 2023 state that: "An Application may claim conformance to the Basic Application Level Confidentiality Profile and Options as a deidentified if it protects and retains all Attributes as specified in the Profile and Options." (Committee, 2023, De-identifier section). The basis for deidentification efforts is formed by the basic profile described in PS3.15 2023b E.1-1. Prior to rollout, deidentification options defined in the supplement can be specified during configuration. For example the option "Retain Patient Characteristics Option" defined therein can be selected to retain information like the patient's ethnicity. After configuration, a human-readable deidentification recipe defines how different DICOM tags are handled (cf. Figure 2). For instance, a patient's name and contact information as well as the physician's information need to be redacted to prevent identification of the patient. In theory, any DICOM tag can be removed or modified with the exception of the ones defined as mandatory by the DICOM standard. This step is responsible for implementing the mandatory guidelines on data protection defined in a data management plan or stakeholder agreement. Other data such as technical configurations of e.g. a CT scan may be preserved to allow for downstream analysis. This is accomplished through the use of pydicom's deidentification framework (Sochat et al., 2018). The same principle can be applied to any data format containing PHI. Generally, LSH supports two forms of metadata modification: whitelisting and blacklisting. Whitelisting assumes that all metadata from DICOM data is removed except for the ones defined on the whitelist. Blacklisting inverts that concept; all metadata is allowed except for the tags defined on the blacklist. Generally, we advise using the more restrictive whitelisting approach, because it is easier to review which metadata is used.

4.7. Git Centered Development and Operations

For LSH, we use a declarative workflow to deploy instances for dedicated studies. Each study is defined by its own Git repository. This is derived from the source repository via a so-called fork. This allows project-specific configurations to be made in the fork on a project-by-project basis. At the same time, changes to the code base can be reflected through the link in the forked LSH instances that have already been rolled out, so that fixes in study-related projects can be transferred back to the source repository. Through this workflow, Git becomes a central place for managing and maintaining all LSH instances. During the rollout or continuous delivery, we leverage the *Kubernetes* integration of *GitLab*, which is also directly integrated into the Web UI. This happens based on individual projects. After the repository is connected to the Kubernetes cluster, the required Kubernetes applications are rolled out for each project. Part of this is a runner that manages and executes the Continuous Integration / Continuous Delivery (CI/CD) pipelines. This principle ensures that the essential issues of separation of duties and data management are addressed by completely isolating the runtimes.

```
FORMAT dicom
%header
ADD PatientIdentityRemoved YES
# Curve Data"(50xx,xxxx)"
REMOVE contains:^50.{6}$
# Overlay comments and data
REMOVE contains:^60.{2}[34]000$
# Private tags ggggeeee
# where gggg is odd
REMOVE contains:^.{3}[13579].{4}$
# Change tags by explicit name
BLANK AccessionNumber
```

Figure 2. Excerpt of a deidentification recipe

5. Use Cases

LSH was developed because of a growing demand for a suitable research study management solution. It is currently in trial use in several projects researching various forms of cancer. The following use cases represent the initially most requested ones and how they shape LSH's design.

5.1. Use Case: Data Collection, Conversion, and Harmonization

Data collection stands at the core of any medical research study. LSH provides an easy-to-use web interface that allows users to register patients and upload data associated with them quickly. Researchers and clinicians can easily share data collected at multiple sites. All DICOM data can be processed using the same pipeline. This ensures that all data has a common entry point and remains accessible for DICOM parsers like *pydicom* (Mason, 2011). Custom pipelines

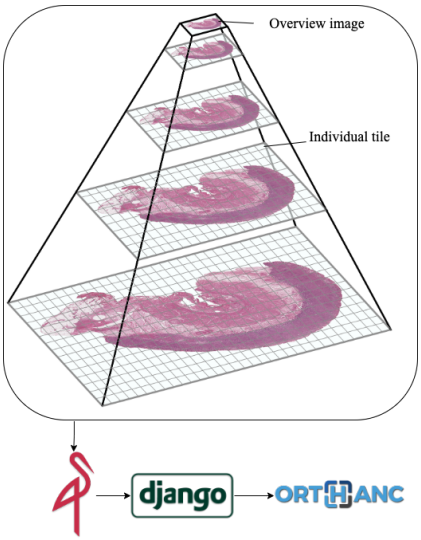

Figure 3. Pipeline for whole-slide image conversion

can be designed to handle frequently encountered, non-DICOM file formats. One such example are histopathological whole-slide images. These files can be notoriously large in size (up to 100 GB per image) because multiple, tiled zoom layers can be contained in the file format (cf. Figure 3). In order to make them more manageable in size, they can be converted to DICOM images. For this purpose, different implementations can be used, such as *dicom_wsi* developed by Gu et al., 2021. However, Orthanc also implements this functionality in its own *OrthancWSIDicomizer* that works well with the Orthanc PACS. Based on these tools, a custom pipeline (cf. Figure 3) is constructed to convert this data to a more manageable file format and to make it accessible from the PACS.

5.2. Use Case: Viewing and Sharing Data

The deployment of a dedicated PACS provides clinicians and other personnel with a centralized, familiar place to view and share (imaging) data. Viewers such as OHIF, or *Stone Web Viewer* (Jodogne, 2022) are able to visualize images from different modalities. Users from different sites can thus easily cooperate in analyzing and discussing patient data. Users can draw from the rich toolset OHIF provides, such as taking measurements, making annotations, or downloading images for other use. Converted histopathological data may require a different viewer due to the pyramidal, tiled structure of this data. Orthanc can make use of *OpenLayers*, an open-source viewer that is able to visualize the converted file properly.

5.3. Use Case: Providing an Infrastructure basis for Application Development

As a study progresses, new feature requests may come up. These can be fulfilled by extending the existing deployment with new applications or prototypes. Although the argument can be made that any existing solution can be extended with new functionality, this process is strikingly straightforward in LSH as it has been a reoccurring topic. We firmly believe that the technical constraints of static installations must not limit a study: Instead of participants having to adjust their workflows to be compatible with an existing deployment, it should be flexible enough to fit changing needs. This is reflected by numerous design choices, such as the microservice-architecture, containerization, and CI/CD. A new microservice interacting with existing ones can thus be developed and prototyped in a secluded container before being rolled out to a deployment using the predefined CI/CD workflow. Services like the FHIR server and the Orthanc PACS possess RESTful APIs that facilitate data exchange, which new microservices can build on top of.

Training segmentation and classification models or using existing ones for inference requires direct access to data. To generate training data, e.g., *Label Studio* (Tkachenko et al., 2020) can be deployed in LSH to enable users to generate ground truths for training. In this scenario, the data is directly accessed from LSH's S3 buckets to be annotated in Label Studio before being stored. The API can be accessed using various popular programming languages. Alternatively, the annotation functionality of OHIF can be used as well. This means that data can be directly loaded from the S3 instance in, eliminating the need for creating and synchronizing local copies. This solution targets research teams that prefer training models locally or on designated machines. Teams with access to suitable hardware can use *Kubeflow* to leverage the potential of machine learning models directly on Kubernetes. Kubeflow enables the training, serving, and management of models that can be directly trained on data collected during the study stored in the MinIO buckets.

5.4. Testing and Validation

Components such as deidentification, whole slide image conversion, FHIR resource CRUD, etc. were thoroughly tested using dummy data and test users before they were moved to clinical trial use. In addition to pure technical testing, services are deployed in experimental, clinical use as early as possible to test their robustness in real-world, practical settings.

6. Requirement Compliance

In this section, common requirements regarding security, legal guidelines and deployment are covered.

Security and Access Management All confidential data must be secured against unauthorized access. To ensure that only registered and authorized users can access the LSH ecosystem, Teleport and KeyCloak are used as reliable authentication and authorization providers.

Legal Conformity LSH works without external dependencies or services, ensuring that data is only ever transported from the user to storage and vice versa. Moreover, this provides complete control and governance over the collected data. The combination of transparent PHI handling and on-premises hosting makes PHI handling straightforward to understand and regulate. In addition, any operations that create or modify existing data are logged in a file that allows for precise traceability of data from its creation to its current state.

Deployment The adoption of a Git-based, continuous rollout process ensures utmost transparency throughout the entire development lifecycle. By centrally managing the code base, the system achieves maintainability and upholds a commendable level of code quality. Leveraging deployment on a Kubernetes cluster enables the realization of a highly reliable setup, satisfying all necessary requirements. Additionally, the administrative burden is greatly minimized thanks to a substantial degree of automation.

7. Limitations

The storage, use and management of health data is subject to ongoing legal and ethical debates that can only be addressed with technical solutions to a limited extent. LSH still requires stakeholder agreements and data management plans to guide the configuration on a technical level. It cannot be deployed without any legal and organizational considerations, but does streamline the process of implementing a consensus reached prior on a technical level. This is done by means of deidentification, controlling CRUD privileges of participants for microservices, authentication, encryption of data in storage and event logging. Moreover, it does cater primarily to image-based studies. The core building blocks and underlying technical workflows are tailored towards image-based data formats: If non-image based data forms the majority of data collected in a study, other solutions may be more advantageous.

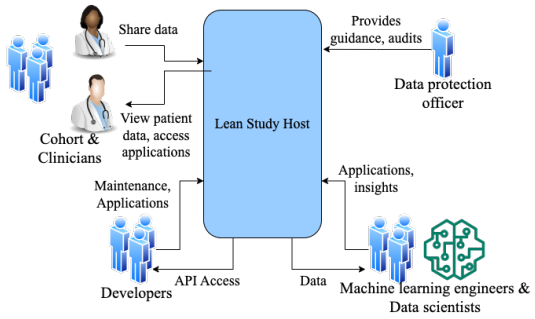

Figure 4. Integration of Lean Study Host into a research routine

8. Conclusion & Future Work

LSH is a modular and easily extensible tool for conducting medical studies, especially in the clinical context. The ability to handle complex, multimodal medical imaging studies makes it relevant for studies in fields such as oncology, neurology, and cardiology. LSH is built around three core user groups (cf. Figure 4).
Clinicians: Experience gathered in previous studies has exposed the need for an easy-to-use, easy-to-understand, and intuitive way of uploading and sharing collected data. Adding data to an existing study is meant to introduce as little interference to the clinical routine as possible. Sharing data securely and functionally is done with the help of a PACS. Tedious efforts such as copying, removing PHI, and sending data are made obsolete by introducing a smart, secure web service that serves as a unified entry point for study data.
Data protection personnel: The proposed solution complies with current legal demands by providing access logs, data deidentification prior to storage, and the option of on-premises hosting.
Researchers and developers: This user group benefits from easy access to raw data and APIs to retrieve data for visualization, model training, and development purposes. Moreover, existing tools and infrastructure can be connected to LSH's ecosystem to bring the comfort of known environments to study practitioners. We summarize our contributions as follows:

- A cloud-native deployment solution for hosting medical, image-based studies in the form of standalone, independent deployments.
- A fully integrated Git workflow, allowing for rapid deployments and adjustments at scale.
- A solution that is highly customizable due to interchangeable modules for all components.
- Adaptable pipelines for harmonization and deidentification of common data formats.
- A privacy-preserving solution that enables fine-grained privilege management to ensure that institutions stay in full control of their data.

The integration of newly arising software standards (such as FHIR) makes it fit for future use under new legislative guidelines. It will thus find continued deployment for all image-based study needs at our institution and be further maintained to provide its services to an increasing number of research groups. On the technological layer of data management and governance, we present a novel ecosystem that lowers the technical overhead of study data management that is packaged in an easy to understand and easy to modify, open-source framework. In future work, we plan to extend the current tool stack with new microservices and features. Moreover we will outline how our approach has impacted the studies that are currently ongoing and document the insights gained from clinical practice.

Acknowledgements

This work received funding from "NUM 2.0".

References

Anastasopoulos, C., Reisert, M., & Kellner, E. (2017). "nora imaging": A web-based platform for medical imaging. *Neuropediatrics*, *48*(S 01), P26.

Bender, D., & Sartipi, K. (2013). Hl7 fhir: An agile and restful approach to healthcare information exchange. *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, 326–331.

Beyene, M., Toussaint, P. A., Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2022). A scoping review of distributed ledger technology in genomics: Thematic analysis and directions for future research. *Journal of the American Medical Informatics Association*, *29*(8), 1433–1444.

Bierer, B. E., Li, R., Barnes, M., & Sim, I. (2016). A global, neutral platform for sharing trial data [PMID: 27168194]. *New England Journal of Medicine*, *374*(25), 2411–2413. https://doi.org/10.1056/NEJMp1605348

Choudhury, A., van Soest, J., Nayak, S., & Dekker, A. (2020). Personal health train on fhir: A privacy preserving federated approach for analyzing fair data in healthcare. *Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30-31, 2020, Proceedings, Part I 2*, 85–95.

Committee, D. S. (2023). Dicom ps3.15 2023b - security and system management profiles, supplement e: Attribute confidentiality profiles. Retrieved September 13, 2023, from https://dicom.nema.org/medical/dicom/current/output/html/part15.html#sect_E.1

Crawford, K. L., Neu, S. C., & Toga, A. W. (2016). The image and data archive at the laboratory of neuro imaging [Sharing the wealth: Brain Imaging Repositories in 2015]. *NeuroImage*, *124*, 1080–1083. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.04.067

Davidson, E., Wessel, L., Winter, J. S., & Winter, S. (2023). Future directions for scholarship on data governance, digital innovation, and grand challenges. *Information and Organization*, *33*(1), 100454. https://doi.org/https://doi.org/10.1016/j.infoandorg.2023.100454

Fayad, Z., & Fuster, V. (2000). Characterization of atherosclerotic plaques by magnetic resonance imaging. *Annals of the New York Academy of Sciences*, *902*(1), 173–186.

Fleagle, S., Johnson, M., Wilbricht, C., Skorton, D., Wilson, R., White, C., Marcus, M., & Collins, S. (1989). Automated analysis of coronary arterial morphology in cineangiograms: Geometric and physiologic validation in humans. *IEEE Transactions on Medical Imaging*, *8*(4), 387–400. https://doi.org/10.1109/42.41492

Grafenstein, M. (2022). Reconciling conflicting interests in data through data governance. an analytical framework (and a brief discussion of the data governance act draft, the data act draft, the ai regulation draft, as well as the gdpr). *HIIG Discussion Paper Series No. 2022-02*. https://doi.org/http://dx.doi.org/10.2139/ssrn.4104502

Gu, Q., Prodduturi, N., Jiang, J., Flotte, T. J., & Hart, S. N. (2021). Dicom_wsi: A python implementation for converting whole-slide images to digital imaging and communications in medicine compliant files. *Journal of Pathology Informatics*, *12*(1), 21. https://doi.org/https://doi.org/10.4103/jpi.jpi_88_20

Gubser, R. J., Schulte-Althoff, M., Heinemann, N., Pohle, J., & Fürstenau, D. (2023). Data governance strategies for data platforms–a multiple case study in nursing care. *ECIS 2023 Research-in-Progress Papers. 50.* https://aisel.aisnet.org/ecis2023_rip/50

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (redcap)—a metadata-driven methodology and workflow

process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/https://doi.org/10.1016/j.jbi.2008.08.010

Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., & Marcus, D. S. (2016). Xnat central: Open sourcing imaging research data [Sharing the wealth: Brain Imaging Repositories in 2015]. *NeuroImage*, *124*, 1093–1096. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.06.076

Hörst, F., Ting, S., Liffers, S.-T., Pomykala, K. L., Steiger, K., Albertsmeier, M., Angele, M. K., Lorenzen, S., Quante, M., Weichert, W., et al. (2023). Histology-based prediction of therapy response to neoadjuvant chemotherapy for esophageal and esophagogastric junction adenocarcinomas using deep learning. *JCO Clinical Cancer Informatics*, *7*, e2300038.

Jodogne, S. (2018). The orthanc ecosystem for medical imaging. *Journal of digital imaging*, *31*, 341–352.

Jodogne, S. (2022). Rendering medical images using WebAssembly. *Proc. of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies*, *2*, 43–51. https://doi.org/10.5220/0000156300003123

Lewin, M., Poujol-Robert, A., Boëlle, P.-Y., Wendum, D., Lasnier, E., Viallon, M., Guéchot, J., Hoeffel, C., Arrivé, L., Tubiana, J.-M., et al. (2007). Diffusion-weighted magnetic resonance imaging for the assessment of fibrosis in chronic hepatitis c. *Hepatology*, *46*(3), 658–665.

Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, *23*(5), 899–908. https://doi.org/10.1093/jamia/ocv189

Marcus, J. S., Martens, B., Carugati, C., Bucher, A., & Godlovitch, I. (2022). The european health data space. *IPOL — Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament Policy Department studies, 2022*. https://doi.org/http://dx.doi.org/10.2139/ssrn.4300393

Mason, D. (2011). Su-e-t-33: Pydicom: An open source dicom library. *Medical Physics*, *38*(6Part10), 3493–3493.

Patel, U. B., Taylor, F., Blomqvist, L., George, C., Evans, H., Tekkis, P., Quirke, P., Sebag-Montefiore, D., Moran, B., Heald, R., et al. (2011). Magnetic resonance imaging–detected tumor response for locally advanced rectal cancer predicts survival outcomes: Mercury experience. *Journal of Clinical Oncology*, *29*(28), 3753–3760.

Scherer, J., Nolden, M., Kleesiek, J., Metzger, J., Kades, K., Schneider, V., Bach, M., Sedlaczek, O., Bucher, A. M., Vogl, T. J., et al. (2020). Joint imaging platform for federated clinical data analytics. *JCO clinical cancer informatics*, *4*, 1027–1038.

Shaban-Nejad, A., Lavigne, M., Okhmatovskaia, A., & Buckeridge, D. L. (2017). Pophr: A knowledge-based platform to support integration, analysis, and visualization of population health data. *Annals of the New York Academy of Sciences*, *1387*(1), 44–53.

Sochat, V., Kolowitz, B. J., Zetterberg, P. K., Ulén, J., & kolowitzbj. (2018, September). *Pydicom/deid: Deid version 0.1.16* (Version 0.1.16). Zenodo. https://doi.org/10.5281/zenodo.1410461

Sun, C., Xu, A., Liu, D., Xiong, Z., Zhao, F., & Ding, W. (2020). Deep learning-based classification of liver cancer histopathology images using only global labels. *IEEE Journal of Biomedical and Health Informatics*, *24*(6), 1643–1651. https://doi.org/10.1109/JBHI.2019.2949837

Tkachenko, M., Malyuk, M., Shevchenko, N., Holmanyuk, A., & Liubimov, N. (2020). Label studio: Data labeling software. Retrieved September 13, 2023, from https://github.com/heartexlabs/label-studio

Watson, C., et al. (2022). Many researchers say they'll share data—but don't. *Nature*, *606*(7916), 853–853.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Ziegler, E., Urban, T., Brown, D., Petts, J., Pieper, S. D., Lewis, R., Hafey, C., & Harris, G. J. (2020). Open health imaging foundation viewer: An extensible open-source framework for building web-based imaging applications to support cancer research. *JCO Clinical Cancer Informatics*, *4*, 336–345.