applied sciences

MDPI

Article

The Holistic Perspective of the INCISIVE Project—Artificial Intelligence in Screening Mammography

Ivan Lazic [1], Ferran Agullo [2], Susanna Ausso [3], Bruno Alves [4], Caroline Barelle [4], Josep Ll. Berral [2], Paschalis Bizopoulos [5], Oana Bunduc [6,*], Ioanna Chouvarda [7], Didier Dominguez [3], Dimitrios Filos [7], Alberto Gutierrez-Torre [2], Iman Hesso [8], Nikša Jakovljević [1], Reem Kayyali [8], Magdalena Kogut-Czarkowska [9], Alexandra Kosvyra [7], Antonios Lalas [5], Maria Lavdaniti [10,11], Tatjana Loncar-Turukalo [1], Sara Martinez-Alabart [3], Nassos Michas [4,12], Shereen Nabhani-Gebara [8], Andreas Raptopoulos [6], Yiannis Roussakis [13], Evangelia Stalika [7,11], Chrysostomos Symvoulidis [6,14], Olga Tsave [7], Konstantinos Votis [5] and Andreas Charalambous [15]

[1] Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Serbia
[2] Barcelona Supercomputing Center, 08034 Barcelona, Spain
[3] Fundació TIC Salut Social, Ministry of Health of Catalonia, 08005 Barcelona, Spain
[4] European Dynamics, 1466 Luxembourg, Luxembourg
[5] Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece
[6] Telesto IoT Solutions, London N7 7PX, UK
[7] School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[8] Department of Pharmacy, Kingston University London, London KT1 2EE, UK
[9] Timelex BV/SRL, 1000 Brussels, Belgium
[10] Nursing Department, International Hellenic University, 57400 Thessaloniki, Greece
[11] Hellenic Cancer Society, 11521 Athens, Greece
[12] European Dynamics, 15124 Athens, Greece
[13] German Oncology Center, Department of Medical Physics, Limassol 4108, Cyprus
[14] Department of Digital Systems, University of Piraeus, 18534 Piraeus, Greece
[15] Department of Nursing, Cyprus University of Technology, Limassol 3036, Cyprus
* Correspondence: oana@thridium.com

check for updates

Citation: Lazic, I.; Agullo, F.; Ausso, S.; Alves, B.; Barelle, C.; Berral, J.L.; Bizopoulos, P.; Bunduc, O.; Chouvarda, I.; Dominguez, D.; et al. The Holistic Perspective of the INCISIVE Project—Artificial Intelligence in Screening Mammography. *Appl. Sci.* **2022**, *12*, 8755. https://doi.org/10.3390/app12178755

Academic Editors: Dermatas Evangelos, Athanasios Koutras, Ioanna Christoyianni and George Apostolopoulos

Received: 20 July 2022
Accepted: 26 August 2022
Published: 31 August 2022

Abstract: Finding new ways to cost-effectively facilitate population screening and improve cancer diagnoses at an early stage supported by data-driven AI models provides unprecedented opportunities to reduce cancer related mortality. This work presents the INCISIVE project initiative towards enhancing AI solutions for health imaging by unifying, harmonizing, and securely sharing scattered cancer-related data to ensure large datasets which are critically needed to develop and evaluate trustworthy AI models. The adopted solutions of the INCISIVE project have been outlined in terms of data collection, harmonization, data sharing, and federated data storage in compliance with legal, ethical, and FAIR principles. Experiences and examples feature breast cancer data integration and mammography collection, indicating the current progress, challenges, and future directions.

Keywords: medical images; mammography; artificial intelligence; deep learning; health data sharing

1. Introduction

Cancer offers a unique context for medical decisions, given not only due to its variegated forms with the evolution of the disease, but also regarding the need to consider the individual condition of patients, their ability to receive treatment, and their responses to treatment [1]. Despite improved technologies, challenges remain in the accurate and early detection, tumor classification/characterization, the prediction of tumor evolution (either locally, recurrently, or metastatically), and the precise evaluation of treatment schemas and follow-up monitoring of cancer.

Medical imaging is an important part of cancer protocols applied mainly (but not limited) to diagnosis and detection stages, providing a variety of information on the tumor's exact location, structure, metabolism, and functions [2]. The main advantages of medical imaging techniques include monitoring capability in real time, minimally invasive procedures, early detection (even for asymptomatic patients), accurate information on the tumor's precise location, as well as details on its size and morphology. Most importantly, they support healthcare providers in the definition of the treatment plan, the evaluation of its effectiveness, as well as follow-up interventions.

The increasing amount and availability of collected data (including cancer imaging) and the development of novel technological tools based on artificial intelligence (AI) and machine learning (ML) provide unprecedented opportunities for improved cancer detection and classification [3], tumor segmentation [4], image optimization, radiation reduction [5], and clinical workflow enhancement. The current imaging methods may be enhanced by identifying findings, whether detectable or not by the human eye, moving from a subjective perceptual skill to a more objective one.

So far, on the one hand, an already-huge amount of cancer data, including imaging data, increases every day. However, the datasets are scattered all around the world and are not used efficiently. On the other hand, AI and machine learning techniques that can provide efficient cancer disease management pathways already exist. However, there is a lack of publicly available AI-ready imaging data, and there is poor confidence in AI-based technology among healthcare professionals for disease diagnosis, prediction, and follow-up [6].

In order to address the challenges related to AI in health imaging, the European commission had issued a tailored call and funded four research and innovation actions in 2020. INCISIVE (a multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction, and follow-up of cancer, no 952179), as one of them, brings together 27 partners from 9 countries (Belgium, Cyprus, Finland, Greece, Italy, Luxemburg, Serbia, Spain, and the UK), aiming to enhance cancer diagnosis and prediction using AI and Big Data in four cancer cases: breast, lung, colorectal, and prostate cancer. In the context of the INCISIVE project, a wide variety of existing cancer images, e.g., computerized tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), ultrasounds (USs), X-rays, and mammograms (MGs), are exploited and further combined with other available data sources, such as biological, medical history, and clinical analysis, offering a critical mass of available data in the envisaged repository of medical images, enabling the excessive training of foreseen AI models. The final outcome was envisioned as a platform offering an AI-based toolbox and an interoperable federated pan-European data repository with secure data sharing and data donorship schema, while at the same time being engineered to meet user needs and improve user acceptance. To achieve a set of ambitious objectives, multiple legal and technological challenges had to be addressed in relation to privacy and ethical considerations, data collection, data integration, de-identification, harmonization, federated data storage, and federated learning.

This article outlines how these challenges are approached in building the INCISIVE solution. We share experiences in tailoring the legal framework appropriate for both EU and non-EU partners; the adopted technological solutions for privacy-preserved data collection, integration, and harmonization; and how federated data storage and sharing has been achieved. These legal and technological grounds of INCISIVE platform are further elaborated, taking the example of breast cancer and mammography as a basic cost-effective imaging modality for breast cancer screening and follow-up. An insight into the process of data collection, selection, integration, and harmonization has been provided, which involves the sharing of actual experiences and obstacles in the process.

The paper is organized as follows. Section 2 describes the challenges towards the use of AI in the medical domain in detail, including the legal and ethical issues that need to be considered. The rest of the text describes the INCISIVE approach implemented in response to these challenges. Section 3 provides an in-depth description of the data preparation and integration tools required before sharing data through the INCISIVE platform. Section 4 depicts the federated data storing approach in INCISIVE, enabling the compliance of legal, ethical, and FAIR principles, while Section 5 describes the AI toolbox implemented within the INCISIVE to optimize breast cancer disease management pathways. Finally, Section 6 concludes the paper and highlights the next steps of INCISIVE development based on the current research.

2. Relevance and Challenges

2.1. Bring Together AI and Medical Experts

AI has a long tradition in computer science and is gradually changing the landscape of healthcare and biomedical research. The boost that AI offers the medical domain includes the enhancement of clinical diagnosis and decision-making performance, such as support for diagnosing patients, as well as making prognostic prediction or selecting treatments through clinical decision support (CDS) systems [7]. The contribution of AI in the oncology field is notable, due to the use of medical imaging and digital pathology outcomes for patients' stratification and cancer prevention which offer further indications for decode molecular signaling cascade and cancer mechanism [8].

Additionally, AI improves the care of chronic disease patients, suggesting precision therapies for illnesses with multiple complications, reducing medical errors [7,9]. However, it is worth noting that AI is not fully autonomous and cannot override human involvement. In the above context, healthcare professionals have to serve as knowledgeable and supportive guides and leaders in this process, making correct diagnoses at the highest possible level [10,11].

However, this venture includes complex challenges that both sides—AI developers and medical experts—have to face properly. From the AI developers' point of view, complex challenges, particularly in the integration, fusion, and mapping of various heterogenous data, still remain. On the other side, clinicians will need to adapt to their new roles as information integrators, interpreters, and patient supporters, and the medical education system will have to provide them with the tools and methods to do so.

Prior to this, several studies highlighted [12,13] many of the issues that are brought up by clinicians which are relevant to any tool or instruments introduced into a medical context, including the system's accuracy across representative cases, as well as the medical point-of-view represented by the AI assistant (e.g., due to choices in training data or labeling). Tools that came up from the above procedure have to be evaluated from a medical and commercial point of view as well. The main question surrounding who will end up controlling, certifying, or profiting from the application of AI still remains, and therefore there is uncertainty around the balance of regulatory safeguards and market forces to ensure that patients who benefit most must be a high priority.

To date, related existing research and innovation initiatives are only limited to small-scale demonstrations, without being excessively validated for reproducibility and generalizability, and without exploring large datasets [10]. In order to explore the full potential of AI solutions in cancer imaging, legal and ethical issues are open challenges which should be addressed in the context of the INCISIVE research project.

2.2. Legal, Ethics, and Data Sharing

The ethical and legal challenges of artificial-intelligence-driven healthcare are well documented in the literature [14]. Hence, it is crucial to take these aspects into consideration. In response to this, the INCISIVE consortium has committed to conducting all research activities within the project in compliance with: (1) ethical principles (including the highest standards of research integrity) and (2) applicable international, EU, and national laws. Hence, all relevant ethical approval actions were acquired in advance before conducting relevant research activities within the project. The proper planning and monitoring of ethical approval actions is of paramount importance during collaborative EU projects. In the current project, a pro-active approach was adopted during the very early stages of the project by starting relevant ethical approval actions, from the second month of the project and initiating discussions across the different clinical sites to enquire about the specific ethical processes and requirements. This early start was deemed necessary when having multiple partners from different countries to avoid any potential delays. The monitoring of ethical approval actions across the different clinical sites/partners was achieved through regular and active email communications and follow-ups, in addition to the use of Excel logbooks to keep track of the status of these ethical applications across different sites. Proper documentation was also deemed crucial; all ethical approval copies alongside English translations were collected from different clinical sites to keep track of these records, as per the INCISIVE Grant Agreement. Several ethical approval actions have been acquired to date. The first aspect of ethical approval actions was related to research conducted to define INCISIVE user requirements across clinical sites. The second aspect of ethical approval actions was related to the retrospective and prospective data collection of health and medical data for the development and training of AI models and data analytics, the testing of AI services, and the development of the INCISIVE repository. The third aspect of ethical approval actions will be related to INCISIVE evaluation studies.

Legal Challenges

Achieving the project's goals presents many novel legal challenges; hence, privacy and ethical considerations were necessary during the early stages of the INCISIVE lifecycle. Alignment with data protection requirements, GDPR in particular [15], is one of the baseline requirements and pivotal points which can guide the development of the incisive repository and toolbox.

In order to develop AI tools, it is crucial to gather medical data from participating health organizations (hospitals acting as data providers). Difficulties resulting from the fragmented legal landscape for the use or re-use of medical data in the EU are widely known [15]. Thus, the first hurdle was to establish the legal basis and conditions for sharing data from each data provider facility, acting on the condition that the data would be processed in a de-identified (pseudonymized) form. In addition to GDPR [15] requirements applied in all member states, national legal provisions which regulate the data providers' ability to obtain retrospective data were investigated. An additional layer of complexity was added by laws applicable to partners from non-EU countries (Serbia). Moreover, data providers' data protection officers (DPOs) were contacted to confirm conformity of the submission of data from their facility with the local hospital rules, which was carried out in parallel with obtaining ethical approval, as mentioned above. The details of each partner's DPOs (or privacy contact) were collected for emergency situations and ongoing compliance assurance. Lastly, ethical considerations [16] were validated throughout the process. Legal and ethical findings and requirements were enclosed in the incisive data management plan, which also specified protocols and tools for the de-identification of shared data.

Next, the foreseen exchange of patient data between the partners' required execution of data-sharing agreements was crucial. Implementing the appropriate arrangements between consortium partners and deciding on the roles of particular organizations in multi-stakeholder research project were both recognized as challenges [17]. In INCISIVE, after the evaluation of the processing circumstances, the drafted data sharing agreements comprised of a joint controllership agreement and two data-processing agreements. The purpose of those agreements was to provide a legal structure for the sharing and use of pseudonymized medical data between data providers and technical partners engaged in AI development tasks. Notably, not all the partners of the consortium were considered controllers; as such, they had no decisions over the purposes and means of processing (such an advisory partners), and were not qualified as co-controllers. For certain parts of the processing, partners providing temporary storage infrastructure or business analysis were considered processors. The drafts of the agreements underwent consultation with all the affected partners. The agreements were planned to be updated if new circumstances arose, such as new partners or changed processing requirements.

In the literature, federated learning is considered a novel privacy-enhancing technique, which can be used to mitigate privacy risks for individual patients [18]. However, the project approaches the complex nature of the sensitive data processing with caution, through the use of new technologies stemming from various sources. The importance of combining both legal and technical protections and the challenges associated with adopting a holistic approach, when it comes to the privacy of biomedical data have been underlined in many studies [19]. Thus, in line with the privacy by design principle, from the early phases of the prototype development, the purpose of processing and potential challenges and limitations needed to be evaluated. To this end, conducting several iterations of the data privacy impact assessment was considered necessary to map and address risks related to the processing of data.

The project's ambition is to develop a legal framework, which enables data providers to share data with other researchers, using federated and hybrid infrastructure. This INCISIVE approach will allow data providers to retain control over their shared data, hosted in their own secure computing environment. The data provider may also use a dedicated central environment for storing their data (hybrid approach). However, adapting the federated approach to the full extent has some trade-offs [20], which need to be considered both from a practical and legal perspective. This will be addressed in the data-sharing legal framework, which will be one of the results of the project.

The goal of the INCISIVE repository is to enable the secure sharing of data in compliance with ethical, legal, and privacy demands, consequently increasing accessibility to datasets for various stakeholders and enabling the experimentation of AI-based solutions. Currently, efforts focus on aligning the requirements of AI researchers, FAIR (findable, available, interoperable, and reusable) principles [21], as well as transparent data sharing and privacy considerations, which are all focal points for the design of the environment. The study takes also into account the alignment of the structure of data sharing with the existing and planned legal acts, including, but not limited to, GDPR [15], the Data Governance Act [22], the Digital Markets Act [23], proposals for the AI Act [24], and European Health Data Space regulations [25]. Those provisions are considered in relation to possible data governance structures, requirements, and processes for platform registration, as well as the use and terms of data sharing. In the process of development and evaluation of the INCISIVE AI-based toolbox, three distinct Study Protocols were needed:

1. A retrospective study—the model training/development of the INCISIVE AI toolbox.
2. A prospective observational study—the validation of the INCISIVE AI toolbox.
3. A prospective feasibility study—the evaluation of the INCISIVE AI toolbox.

Through a rigorous collaborative approach, all aspects of each of the study protocols, including the ethical considerations, have been discussed and agreed upon collaboratively between the relevant consortium partners and, where necessary, the whole consortium. The challenge in relation to the ethical considerations of the three studies stemmed primarily from the diverse processes in place for seeking ethical approval in the five countries where the studies would be deployed. Because of this diversity, it was decided that general ethical template resources would be provided to the stakeholders which can be later modified and adjusted to the specific local requirements and processes in place. Another challenge, in this context, was the diverse group of the target populations. As these include patients, clinicians, and technical personnel, all the appropriate ethical documentation should accommodate the different ethical considerations and hence correspond to their diverse status. The international good practices for the evaluation of technological solutions have informed the decision-making process [16]. The first stage of the development phase included internal meetings to decide on the most appropriate methodology for the study protocols and explore what ethical implications need to be taken into consideration. Things that were taken into consideration in this process included the aim of the study, the type of the intervention to be applied in the clinical setting (e.g., whether the INCISIVE generates ethical concerns in relation to data management), modalities to be assessed, tools to collect the required data (e.g., GDPR principles), ethical issues (e.g., consent forms, withdrawal forms) the type of users, and the variations of the recruitment sites (e.g., recruitment strategy) where the INCISIVE system would be deployed. At this early point of the study protocol development, all progress in the development of the INCISIVE system has been taken into consideration so that it is possible to correlate the two developmental processes (INCISIVE system and study).

As part of the development process, and following a thorough presentation of the partners' thoughts on the design and implementation of the study protocols, various means of data gathering have been implemented to allow the partners to provide their perspectives on specific aspects of the protocol (e.g., data collection, methodological design, recruitment process, etc.) with specific emphasis on the ethical considerations. With regards to the ethical processes and requirements in place in the five participating countries, partners were asked to provide detailed feedback on what information will be necessary to acquire ethical approval. This process was complemented with telephone meetings with specific partners as well as email communications (e.g., to clarify processes). This was deemed necessary to comprehensively capture the specificities of the various recruitment sites and how the developed protocol could best accommodate these.

Following this process, four workshops were held specifically to acquire consensus on certain topics, including the data to be collected as part of the study (including the different modalities per tumor site), the timepoints of collecting the data, human evaluators approach, verifying the clinical sites, and the ethical processes and requirements for the studies. A separate meeting has been held with the recruitment sites to verify the number of cases to be included in each of the studies and to agree on the ethical resources to be provided as templates to partners. Additionally, during the planned plenary meetings, the consortium had the opportunity to collect feedback regarding the planned actions as part of the study protocol definitions.

The next section presents the outcomes of the workshops and provides an in-depth description of the data preparation and integration tools required before sharing data through the INCISIVE platform.

3. Data Preparation and Integration

3.1. Data Integration and Quality Check Tool

During its lifetime, INCISIVE collects data divided in two categories: (a) clinical and biological data in a structured form, including demographic and medical history data, histological and blood markers, and treatment and tumor details, and (b) imaging data, coming from various modalities, including mammography, CT, and MRI. These data correspond to distinct timepoints during the patients' treatment: baseline/diagnosis, after first treatment, first follow-up, or second follow-up.

A multifaceted strategy for data integration was followed which aimed to address multisite data collection challenges as regards non-imaging data and imaging data, including steps for the homogenization of data structures and semantics [26].

Following an iterative procedure to reach a consensus among data providers, a template was created, and data providers were asked to collect the structured data following this template. The data schema is depicted in detail in Figure 1. The template included, among other things, the type of information and the terminology to use in specific fields, specifically adopted for the specific type of cancer. A small subset of fields was considered as mandatory, and their absence was considered to significantly decrease the data usability. The imaging data were used in studies and series per person in a folder hierarchy following various conventions and were accompanied by segmentations of the volumes of interest.

All data providers were given instructions on the steps to follow to make data collection as uniform as possible. In addition, a data quality check was also performed on a local level to detect possible problems and support the data provider in correcting them before uploading data to the INCISIVE repository [27]. These included checks of the structured data (for structural and semantic compliance with the template, the existence of duplicates, and the absence of mandatory values), as well as checks of the imaging data hierarchy and

Digital Imaging and Communications in Medicine (DICOM) files. The data preparation workflow, along with the several components of the data integration quality check tool, is described in Figure 2.

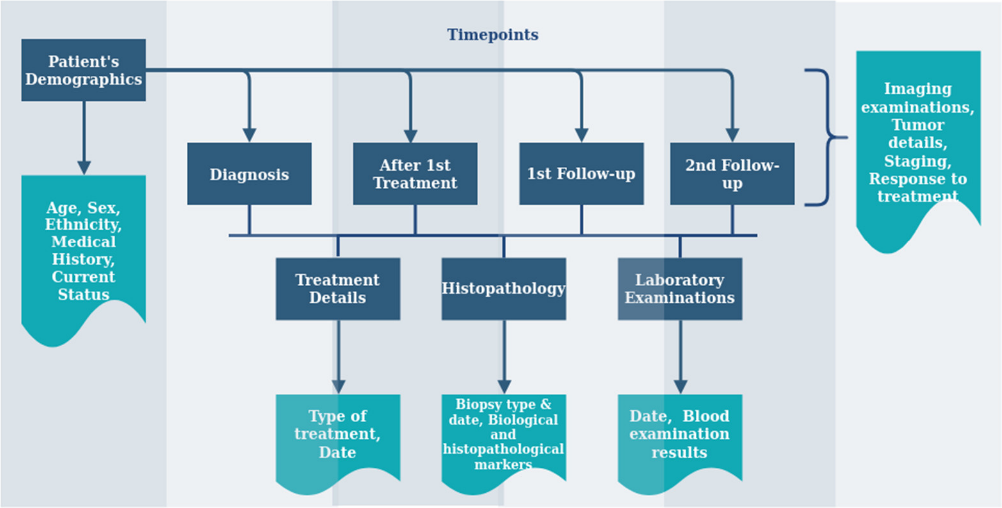

Timepoints
Patient's Demographics
Age, Sex, Ethnicity, Medical History, Current Status
Diagnosis
After 1st Treatment
1st Follow-up
2nd Follow-up
Imaging examinations, Tumor details, Staging, Response to treatment
Treatment Details
Histopathology
Laboratory Examinations
Type of treatment, Date
Biopsy type & date, Biological and histopathological markers
Date, Blood examination results

Figure 1. The data collection schema.

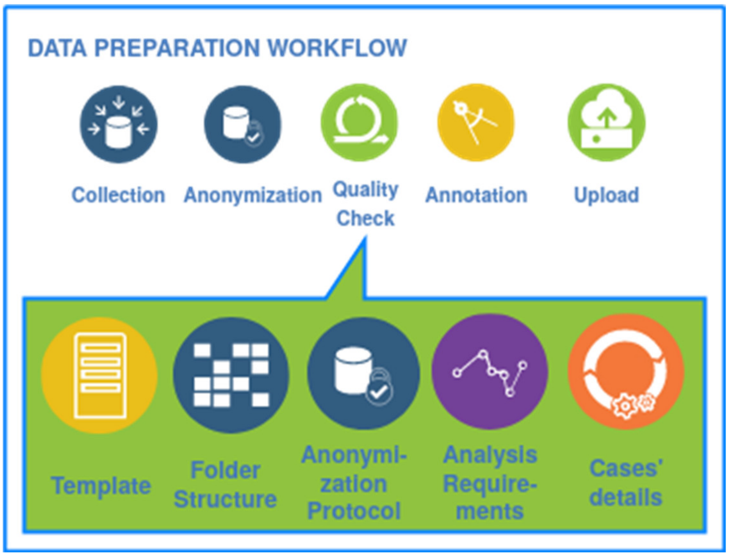


Figure 2. Data preparation workflow.

3.2. Annotation and De-Identification Tools

3.2.1. Annotation

INCISIVE creates a data repository that consists of four different types of cancer (lung, colorectal, breast, and prostate) and six imaging modalities (MRI, PETCT, CT, MG, US, histopathological image, and X-ray), so further actions are required to ensure high quality in the annotation procedure amongst different data providers. The segmentation of organs and tumors, understood as the delineation of these structures, can be divided into three types, mainly depending on the degree of automation of the process and intervention of the clinical user: manual, semi-automatic, and automatic. Three types of annotations were initially planned:

- classification where the whole image is assigned a corresponding label;
- bounding box where a rectangular region of interest is assigned a corresponding label;
- segmentation where each pixel is assigned a corresponding label.

A set of labels was decided cooperatively with all the data providers for each combination of imaging modality and disease type used in INCISIVE.

The tool adopted for the segmentation process was ITK-Snap. The results of the annotation process were exported as a NIFTI file and named using the following information:

- SiteID is completed from each data provider (e.g., 001 for CERTH, 002 for ICCS, etc.).
- PatientID is created by the de-identification procedure, which took place before annotation (e.g., 000001, 000002).
- Modality is completed from each data provider (e.g., MRI, PETCT, CT, MMG, US, HP).
- Timepoint (BL, TP1, TP2, TP3).

The medical expert annotators were instructed to keep their annotation environment as uniform as possible within their organization and within the INCISIVE consortium. This was quite important for the uniformity of the annotation of the data (and therefore their quality) since the INCISIVE project consists of a large number of data providers. Continuous technical support was provided, and guidelines were constructed for the medical expert annotators, especially during the first few months of the project, to ease the process of installation and usage of the annotation tool (ITK-Snap version 3.8.0) [28].

3.2.2. De-Identification

This section describes the de-identification protocol that is currently being used by the INCISIVE data providers, contributing retrospective imaging data to the INCISIVE repository. INCISIVE uses a structured approach based on recognized standards and best practices for the de-identification of DICOM images [29] to ensure that the de-identification process removes or replaces any personal identifier that may lead to patient identification. The INCISIVE de-identification process is developed in compliance with GDPR.

INCISIVE imaging datasets are in the DICOM format, which is a common format that facilitates interoperability between medical imaging devices [27,30]. The DICOM format contains meta-data that often possess identifiable information on the patient, study, institution, etc. Some manufacturers also use private attributes and elements to store complementary information not defined in the DICOM Standard, which is usually undocumented, and could contain patient information in some cases.

Step 1—ID mapping. Data providers are instructed to create an ID mapping table (e.g., excel file), both manually and separately from each other, that consists of two columns:

- The real name of the patient (or the alternative identifiable information that each data provider chooses (e.g., a social security number);
- The ID (a variable number of characters for the pilot's name (e.g., AUTH), followed by a dash, followed by a five-digit incremental number).

Mapping of the original patient ID and the ID that the images have within INCISIVE is performed by each data provider, and no other INCISIVE partner has access to the mapping table. The data provider only communicates the real name ID mapping table within its organization and only the ID within the INCISIVE consortium. The mapping table is stored with the responsibility of each data provider and is defined by each one separately.

Step 2—DICOM de-identification. All submitted data are de-identified locally and separately on each data provider, before they are securely uploaded to the INCISIVE temporary repository. Data providers use the CTP anonymizer [31] as a de-identification tool. The CTP anonymizer is installed on a standard desktop computer to submit sites, and data providers' personnel are trained during the procedure (e.g., dedicated workshop, the provision of instructions on how to use the tool). Data providers have been instructed to not utilize any other de-identification software prior to the INCISIVE system in order to avoid removing any critical information (e.g., possibly a PACS-integrated de-identification tool). In case a data provider chooses to use any other de-identification tool instead of the CTP anonymizer, they then ensure that the tool meets the INCISIVE de-identification profile.

Step 3—The testing of de-identification results. If the result of the test is deemed sufficient further to the profile stated below, the dataset may be uploaded on the tempo-

rary infrastructure. The standard for the de-identification of DICOM objects is defined by the DICOM Standard PS 3.15 [30] that specifies a set of de-identification rules for use in various situations. The rules are grouped into a Basic Application Confidentiality Profile that removes all PHI, while various options are also defined to be applicable to the Basic Application-Level Confidentiality Profile, specifying the removal of additional information or the retention of information that would otherwise be removed.

3.3. Interoperability Standards and Data Model

In INCISIVE, the data (including health and imaging data) transformation to a common data model (CDM) is the cornerstone of the data harmonization process that enables the aggregation of clinical data for data consumers, data findability, data accessibility, and data reusability (in compliance with FAIR principles—findable, available, interoperable, and reusable) [21] for driving the development of trustworthy AI tools that can provide an accurate and prompt diagnosis of cancer, prediction, and follow-up. To date, within the health and life sciences sector, many efforts are being made to standardize vocabularies, establish ontologies, and interoperable standards. Nonetheless, in terms of structural data interoperability, two main interoperability standards (the HL7 FHIR for clinical data and DICOM for medical imaging) are widely used, and probably the most used, in the sector. It is therefore natural that the choice of these two standards has been made for INCISIVE. When it comes to the terminology adopted for semantic interoperability, INCISIVE has based its CDM on the SNOMED ontology for clinical data and LOINC (Logical Observation Identifiers Names and Codes) [32–34] (Figure 3). SNOMED [35] is a standardized, international, and multilingual core set of clinical healthcare terminology, and LOINC is an international standard for identifying health measurements, observations, and documents. Both of them are widely used ontologies, thus ensuring and maximizing the interoperability of INCISIVE data, as well as the sustainability of INCISIVE.

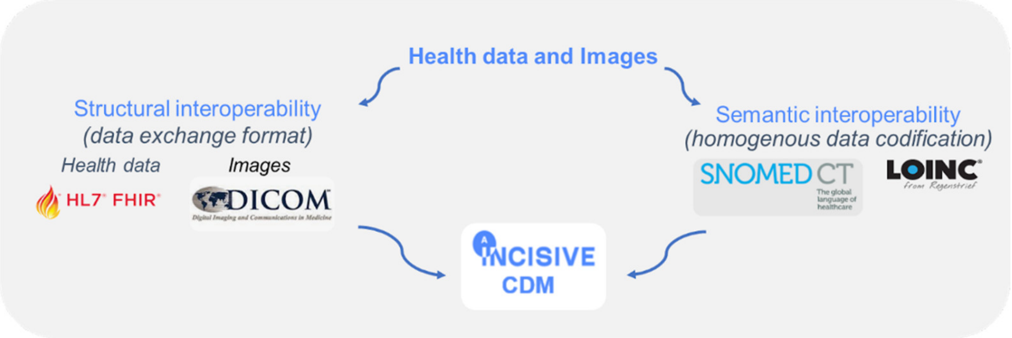


Figure 3. INCISIVE interoperability backbone.

In practice, a list of variables to be collected for each type of cancer addressed in INCISIVE (colorectal, lung, prostate, and breast) was built based on inclusion and exclusion criteria agreed with clinical experts (see Section 3.1). These variables were then mapped to the SNOMED ontology for clinical data and LOINC into a Microsoft Excel template. More than 500 clinical terms were encoded: 177 for breast cancer, 134 for colorectal cancer, 116 for prostate cancer, and 155 for lung cancer.

This Excel template was fetched from an extract transform load (ETL) tool which reads them and generates FHIR messages (xml messages). More than 100 FHIR entries were used to define the four HL7 FHIR messages in the XML format: 38 for breast cancer, 30 for colorectal cancer, 33 for lung cancer, and 29 for prostate cancer. Each FHIR message was sent to an underlying FHIR server that integrates and enables the management of the homogeneous datasets in the INCISIVE infrastructure (see Section 4.2).

In medical imaging, the naturally adopted standard is DICOM, which stands for digital imaging and communications in medicine. More specifically, DICOM is a clinical messaging syntax used to exchange medical images between medical equipment and information systems. This standard enables the distribution, storage, and viewing of medical images. It also allows the integration of scanners, servers, workstations, printers, and network hardware from multiple manufacturers into a picture archiving and communication system (PACS). When it comes to AI development, imaging-related information generated after the acquisition of images, and in particular by radiologists reading exams, is essential and allows the capture of contextual information and the enrichment of data. Beyond DICOM images, DICOM standardizes many other types of data, such as imaging reports, measurements, annotations, and segmentations. For images, INCISIVE follows a similar procedure to clinical data with a second ETL tool that reads DICOM files placed in a directory and then sends them to an underlying PACS server. The PACS server integrates and enables the unified management of DICOM images in the INCISIVE infrastructure, through the esthesis client [36] (see Section 4.2), which is responsible for securely interconnecting the federated nodes and the federated repository. Figure 4 illustrates the INCISIVE ETL process.

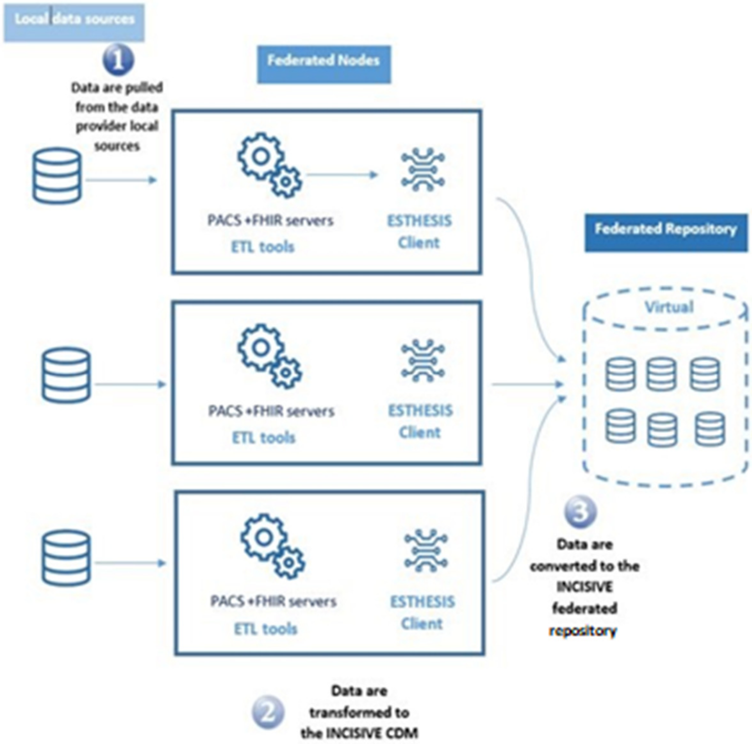


Figure 4. The ETL process in INCISIVE.

In relation to the overall INCISIVE architecture, both ETL-FHIR and ETL-PACS components are provided as Docker containers.

In the future, the INCISIVE common data model is subject to updates in order to integrate all relevant user requirements and specifications. To this end, SNOMED source files may be transformed into the Observational Medical Outcomes Partnership (OMOP) vocabulary structure [37].

3.4. Data Collection and Integration Experiences—Breast Cancer

Once the retrospective data collection study was designed and ethical approval was collected, each data provider could start with a data collection procedure following the provided guidelines related to:

- a common clinical data schema, i.e., a structural embedding designed to suit all types of information;
- standards of medical terminology to be used, as agreed by a consensus among data providers;

- image de-identification (CTP DICOM anonymizer [38]);
- evaluation of the data quality and compliance (Section 3.1);
- image annotation (the ITK-Snap tool [28,39]).

The data organization and naming conventions were imposed to facilitate consistent data structure in the INCISIVE retrospective data repository.

The size and heterogeneity of data in healthcare institutions are vast and depend on a hospital's scope and coverage area. In this project, six institutions from four countries acted as data donors for breast-cancer-related data: the Hellenic Cancer Society (HCS), Greece; the German Oncology Center (GOC), Cyprus; Aristotle University of Thessaloniki (AUTH), Greece; the Vojvodina Institute of Oncology (as a scientific base of University of Novi Sad, UNS), Serbia; Federico II University Hospital (providing an imaging service to the Department of Advanced Biomedical Sciences, University of Naples, DISBA), Italy; and Policlinico Tor Vergata of Rome (Faculty of Medicine of the University of Rome Tor Vergata, UNITOV), Italy. The type and scope of these institutions vary from general hospitals such as the AHEPA University Hospital (AUTH), Federico II University Hospital (DISBA), and Policlinico Tor Vergata of Rome (UNITOV) (a non-profit cancer organization linked to multiple hospitals (HCS) and specialized oncology clinics (GOC and UNS).

Although national guidelines define breast cancer care pathways, they are aligned around internationally recognized good practice in the diagnosis and treatment of breast cancer [40,41]. The care pathways in the participating countries correspond to the clinical workflow diagram presented in Figure 5. The patient journey is usually initiated upon a patient self-examination or as a result of a regular annual screening, and all suspected cases are directed to primary healthcare centers. General practitioners (GPs) collect the information, examine the patient, and ensure that patient has been directed to a hospital (one stop breast clinics) or a specialized center for further examination and diagnoses (including mammography (MG), ultrasound (US), and/or magnetic resonance imaging (MRI) if appropriate). Within the same institution, the patient undergoes fine-needle aspiration (FNA) or core biopsy if appropriate. When the suspected case is confirmed, if needed, the patient undergoes additional imaging, such as liver US, abdominal MR, computerized tomography (CT) scans, bone scans, and positron emission tomography (PET). With these clinical and imaging examinations, the first steps of treatment can be defined. Due to the complexity of the disease, diagnostic and therapeutic tools call for the involvement of many specialists, i.e., radiologists, pathologists, oncologists (clinical and radiation), breast surgeons, and nuclear medicine physicians. In the treatment process, the patient is followed by a multidisciplinary team (MDT), including an oncologist, a radiologist, a breast surgeon, and a pathologist, while the patient care, recovery, and support process is assisted by psychological support and rehabilitation services. Post-treatment monitoring includes MR, US, or/and MR imaging at regular intervals, either annual or bi-annual. The clinical workflow results in multiple time points (TPs) on a patient journey which are documented in the common data schema.

The number of examinations and reports associated with the described clinical workflow is large (Figure 5), and the way this information is stored in the hospital information system is system-dependent and differs among institutions. Information extraction from clinical records and translation to the common data schema requires at least 2 h per patient for an experienced clinician. Difficulties arise as the number of fields is large and additional care has to be taken on the precise timing of the events in the patient timeline. Each patient record is followed by at least one TP, though preferably four or more, in which diagnoses, follow-up, or treatment decisions are documented by an appropriate imaging study.

The whole data collection process at the data provider side is depicted in Figure 6 Besides the de-identification and quality checker tool described in Section 3.1, the annotation of the de-identified imaging studies is carried out according to the predefined INCISIVE guidelines. The annotation is modality-dependent and covers different types of changes, including lesions (malignant or suspicious), calcifications, axilla lymph nodes, suspicious regions, and surgical clips (Table 1). Annotation is performed in the ITK-Snap tool, for each image in the imaging study, implying that numerous images for each study are made using volume-based modalities.

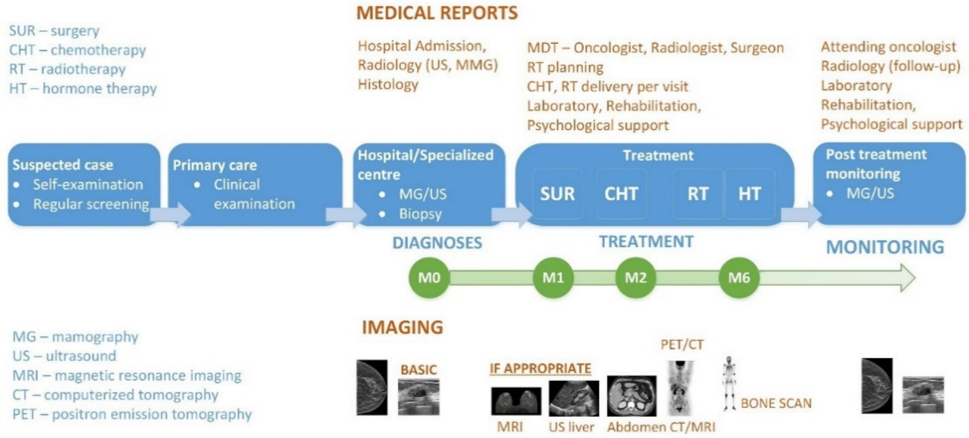


Figure 5. The care pathway in breast cancer with simplified presentation of clinical and imaging data documenting the clinical workflow.

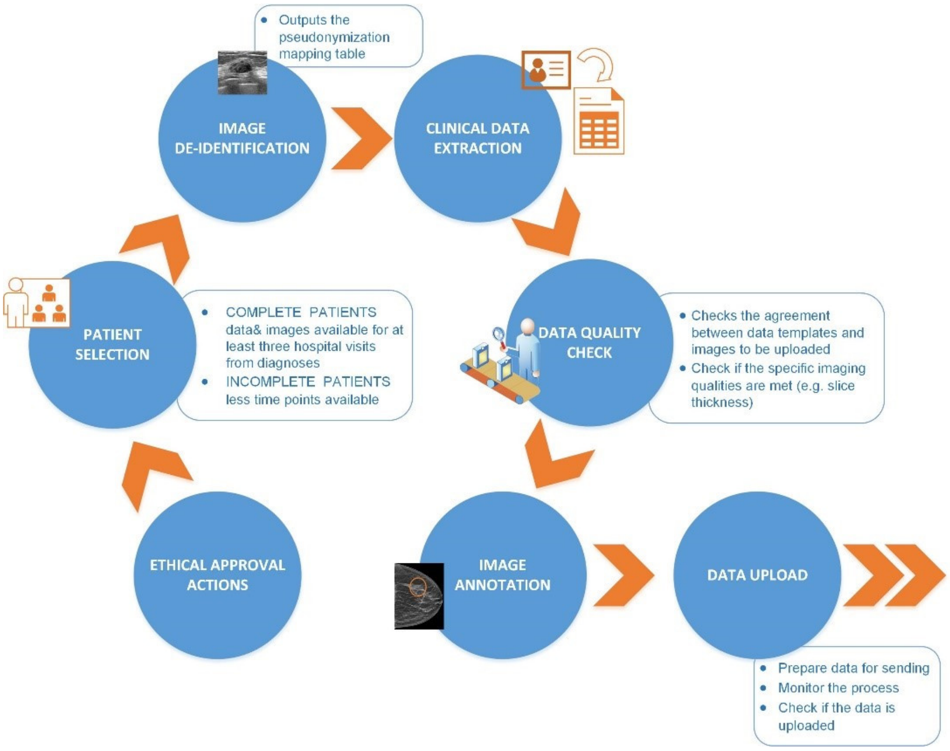

Outputs the pseudonymization mapping table
IMAGE DE-IDENTIFICATION
CLINICAL DATA EXTRACTION
PATIENT SELECTION
COMPLETE PATIENTS data& images available for at least three hospital visits from diagnoses
INCOMPLETE PATIENTS less time points available
DATA QUALITY CHECK
Checks the agreement between data templates and images to be uploaded
Check if the specific imaging qualities are met (e.g. slice thickness)
ETHICAL APPROVAL ACTIONS
IMAGE ANNOTATION
DATA UPLOAD
Prepare data for sending
Monitor the process
Check if the data is uploaded

Figure 6. INCISIVE data collection process.

The final step in the data collection process regards the data upload to the temporary repository. The procedure required user authentication, set-up, and monitoring of the virtual connection for data transmission.

The data collection process was followed by numerous challenges associated with each step in the process. The process of clinical data extraction was time-consuming and error-prone, as the patient timeline had to be strictly followed, hindered by the hospital information system (HIS) design, as medical reports of different departments (e.g., radiology and pathology) are usually clearly separated. For these reasons, all patient specific reports were examined, timelines were restored, and time points were defined before the data input in the clinical data sheets. During the operating hours, the HIS was usually overburdened with requests and data extraction is slowed; for these reasons, data providers with a larger number of patient weekends were the only option for faster data access. Moreover, in cases when the hospital provides for the wider geographical region, some patients were referred to the hospital only for specific interventions, such as surgery, treatment protocol definition, or specific imaging (PET), while other time points on the patient journey were performed locally and information was made unavailable. For these reasons, patients with at least four consecutive TPs available (“complete cases”) were differentiated from those with less or intermittent TPs (“incomplete cases”) in order to facilitate the selection of patients for the training of AI models.

Table 1. The modality annotation convention in the INCISIVE project.

	Ultrasound	Mammography	MRI	CT Scan	PET/CT Scan
Data requirements	Healthy and non-healthy images	• Healthy and non-healthy images • CC and MLO projections • Tomo-synthesis	• MST: 3 mm • Sequences: T2W axial, DWI, T1W, post-contract sequences	MST: 5 mm	MST: 5 mm
Annotation procedure	Bounding box	Contouring	Contouring	Bounding box	Bounding box
Labels	• Benign • Suspicious/ indeterminate • Malignant	• Benign • Suspicious or indeterminate • Malignant • Calcification • Surgical clip • Axilla lymph node	• Benign • Suspicious or indeterminate • Malignant	• Benign • Suspicious or Indeterminate • Malignant • Macrocalcifications	• Benign • Suspicious or indeterminate • Malignant

In cases when a lower number of patients was available for data donation, the whole process, though-time consuming, became manageable. For a larger number of patients, data providers had two strategies: the gradual upload of patients over an extended 2-year period (e.g., HCS), or the development of an expert system for an automated identification protocol and analysis of patient records (e.g., UNS), as described in [42]. The latter approach is demanding, and is language- and HIS-dependent, yet it could serve as a positive example towards the development of a natural language-processing tool as part of an efficient and trustworthy system used to extract information from clinical records.

The data integration quality check tool (as described in Section 3.1) was designed for project needs, and in the development process data providers encountered multiple issues in using and testing this product. The support was continuous, once issues reported were resolved, and the process of checking data quality was repeated from a data provider side. The first software versions did not have intuitive error messages, so users had to resort for help more often. Despite this, the looping process was time-consuming, and multiple feedback ensured thorough software testing, which supported the upgrade and final version of the software for the studies to follow within the project.

DPs identified some flaws in the design of the commercially available ITK-Snap annotation tool, which caused a more time-consuming annotation process. The time leaking during the annotation task was caused by a need to set the loading directory for each DICOM image, as the previous image settings were not preserved. The storage directory for the corresponding annotation may not be set semi-automatically, e.g., the same folder as the loading one, but the last one was preserved and had to be changed on an image basis. The annotation of volumetric images, such as MR or CT scans, requires annotation of multiple images, for which the software offers an automatic interpolation of annotations based on the annotation of several non-consecutive images. However, this option does not work when there are multiple ROIs to be annotated in each image.

4. Results to the Existing Data Repository

4.1. Data Sharing

Typically, technological solutions for data sharing rely on Cloud platforms [43,44]. In INCISIVE, we start from the assumption that the current solutions available on the market are unable to meet the (conflicting) security requirements of privacy-sensitive applications, such as cancer image processing. INCISIVE aims to extend already-existent solutions of data sharing (e.g., NextCloud [45]) and provide data protection measures and GDPR compliant-by-design tools. In INCISIVE, we aim to enhance the privacy-preserving mechanisms by (i) ensuring that even the public–private Cloud operator has no access to the data (data owners can keep their data completely private); (ii) providing a set of metrics useful to measure the degree of privacy of different categories of data; and (iii) ensuring users have the right and freedom of opting out and maintaining full control of their data. INCISIVE adopts a novel approach for ensuring privacy in already-existent data-sharing platforms, relying on security hardware extensions available in commodity CPUs (most notably Intel Software Guard Extensions, SGX [46]). In INCISIVE, operations of ciphering/deciphering and procedures of key management are protected in the trusted execution environment of Intel SGX. INCISIVE can adopt a blockchain-based auditing mechanism for logging all the critical actions associated with data stored on the Cloud, e.g., patient data access via Cloud operators or other unauthorized users. The adoption of blockchain with built-in smart contracts will ensure the tamper-proof decentralized logging, auditing, and tracking of such actions, leading to traceability and non-repudiation. Moreover, the blockchain mechanism will allow for the logging, auditing, and tracking of transactional data activities between the involved INCISIVE layers (data level, analysis level, and user level), e.g., data submission, sharing, and access in a holistic manner that preserves confidentiality and data access permissions.

The impact of such solutions will be tangible via specific privacy metrics, properly designed for the implemented mechanisms. Such metrics are: (i) the attack surface exposed to untrusted software, (ii) the trusted computing base used by security tools in line with source lines of code (SLOC), and (iii) the number of access points to the SGX-isolated environment. With the INCISIVE solution, data owners can avoid trusting the Cloud provider. The threat model covered in this project assumes that even data-sharing platforms could have malicious embedded codes (e.g., backdoors), leading to security flaws.

The system is built upon a federated storage approach and a set of standardized open APIs that will enable the linking of various local data sources (similar to the ones utilized in INCISIVE to the system), interaction with users, communication with processing infrastructures, and the sharing of data. The communications with external systems are based on established standards (HL7 FHIR, DICOM, SNOMED, etc.) and in combination with mapping (e.g., common data models), and the data management functionalities interoperability and interconnectivity are ensured. The system is considered one of the main exploitable outcomes of the INCISIVE project. Its stand-alone nature and interoperability features will enable its connection with third-party trusted AI providers, as well as with established healthcare systems (e.g., PIMED in Catalonia, etc.), contributing to its future sustainability.

The development of new AI methods and applications for image analysis requires large-scale datasets for the training of machine learning algorithms, as classification accuracy is largely dependent on the size and quality of the dataset. Currently, cancer imaging databases are scattered around the EU with no joint coordination across them, being small in size and lacking real-world variation. However, single research centers cannot produce enough data to fit prognostic and predictive models of sufficient accuracy, resulting in algorithms with limited diversity and generalizability to widespread clinical practice. Data sharing in cancer imaging is therefore of utmost importance. Standards, including the PACS and the DICOM, ensure standardized imaging data structure and communication/retrieval. However, most of the imaging datasets are often heavily protected and not accessible for research purposes or accessible imaging data, and are often unusable because they are not appropriately curated, organized, de-identified, or annotated. The imaging data curation involves trained professionals, making the process time- and labor-intensive. Moreover, the training of deep networks requires balanced sample datasets as the inequality of training data distribution can induce biased models. Limited data-sharing initiatives exist, such as the Cancer Imaging Archive (TCIA) and the National Institutes of Health databases among others, whereas the extent of data sharing required for the widespread adoption of AI-based technologies across health systems will require more expansive efforts. Consequently, few appropriate image datasets are available for the research and vendor community to support the development of automated clinical solutions and the limited training data act as a bottleneck for the further application of deep learning (DL) methods in medical image analysis. To address these gaps, large-scale imaging datasets with standardized validation sets that capture a comprehensive diversity of test images (including both routine and challenging cases) will need to be developed for each application in combination with more effective auto- or semi-automated methods for data management standardization to support both clinical practice and research.

One of the major goals of INCISIVE is to address the data availability challenge towards the wide adoption of AI solutions in health imaging. INCISIVE will aggregate and unify the fragmented cancer imaging datasets across European healthcare systems and institutions, characterized by a multiplicity of data sources, enabling the integration and full exploitation of current initiatives and isolated databases to reach a critical mass of gathered data. INCISIVE proposes and implements the pan-European repository of health images following a federated approach, aiming to break the silos by incorporating mechanisms for secure data sharing and ensuring that the data providers keep full control of their data and their right of opting out. The rationale behind this decentralized approach is to minimize centralized storage operations, bringing the INCISIVE features to the local level. Blockchain technology is utilized to achieve transparency and traceability. A data donorship mechanism is incorporated to empower both patients and institutions to contribute imaging data for responsible research, allowing data sources to be easily linked to the system and human quality ratings to be acquired. Compliance to the corresponding legal directives is embodied to the core of the repository, while corresponding standards (HL7 FHIR, DICOM, and SNOMED) are utilized to enhance interoperability and connectivity. Above all, INCISIVE ML-aided data curation, de-identification, and annotation reduce efforts and human interventions relating to these tasks, thus improving the accessibility to isolated datasets even further, making them available for research purposes. The use of data during and beyond INCISIVE is based on FAIR principles to make data discoverable and processable both for human- and machine-driven activities. This will increase their interoperability with current and future imaging cancer datasets, allowing researchers to design and develop AI tools beyond the project's current scope.

The INCISIVE pan-European repository implementation is based on two major axes. The first one concerns the incorporation of the FAIR principles in the development process, in order to ensure that the data meets the following standards: (1) findability, (2) accessibility, (3) interoperability, and (4) reusability. This approach will act as the basis of solving the challenge of data availability. The second one is to ensure that the data providers have full control of their data and simultaneously to promote the advantage of AI applications, striving for the users to build trust on such solutions. We expect that the positive outcomes of the AI-based solutions in the observational and interventional studies of INCISIVE will further reinforce the trust of patients and institutions and motivate data donation.

INCISIVE promotes openly reusable access to de-identified health image datasets across the EU for training AI applications.

The need for access to high-quality datasets, containing appropriate annotations or rich metadata, for developing high-quality AI algorithms is as widely accepted as it is urgent. High-quality de-identified public datasets are very scarce, while interoperability issues, privacy concerns, regulations, and resource requirements further prohibit access and re-useability. Equally vital is access to proper tools for data extraction, de-identification, and labelling; these, however, are very costly.

4.2. Federated Sharing

The improved usage of health data can help to validate the relevance of novel diagnostic and therapeutic methods as well as directly improve care. Patient care may benefit from the availability of actionable data-driven standardized decision support across care-delivering organizations. In that context, INCISIVE aims to become a facilitator for interoperability and interconnection between its stakeholders, and for data as well as services shared for the optimized management of cancer disease. To this end, the INCISIVE pan European federated repository of health images builds on a federated storage approach and employs digital processes and information technology to facilitate interconnection between willing participants, whether data providers and/or data consumers. By leveraging existing data interoperability standards (HL7 FHIR, DICOM, SNOMED, and LOINC), open technology (The Esthesis platform [36]), and concept (federated storage), it enables open, consistent, quality-assured, and easy-to-use innovative data exchange and services for healthcare providers, medical researchers, and AI developers in a federated manner.

The INCISIVE platform is divided into three parts (see Figure 7).

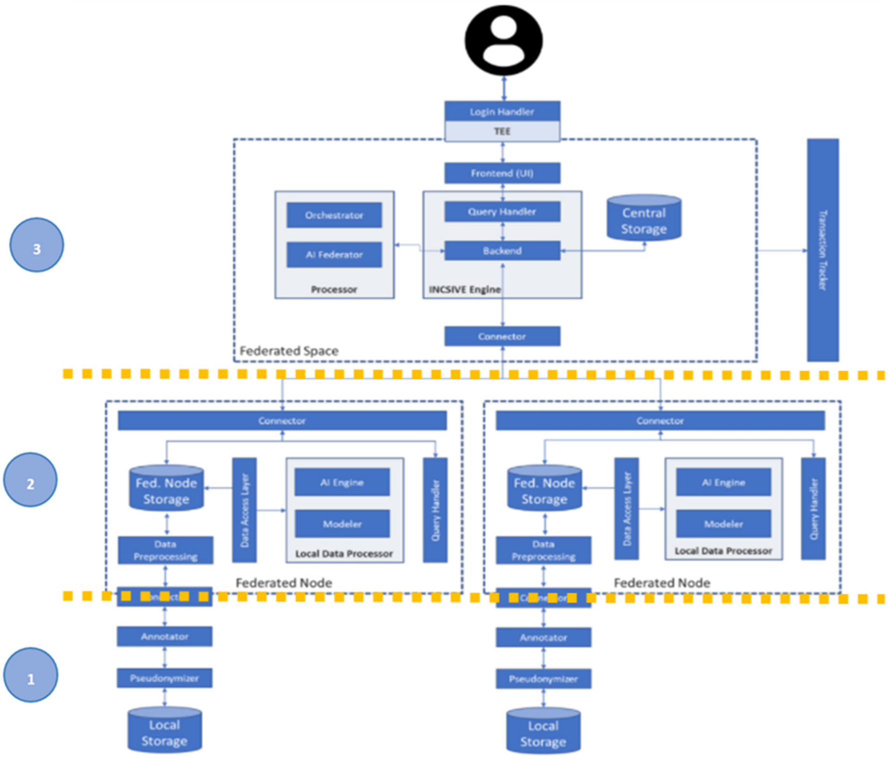


Figure 7. INCISIVE preliminary architecture.

- **The data preparation toolset** is a set of tools that operate at each data provider's site before the data federation stage so that the data becomes GDPR-compliant and appropriately processed (e.g., annotated, where relevant). The main tools that support this process are the data de-identification tool, the data annotation tool, the data curation tool, and the data quality check tool. These tools ensure that the data undergoes a correct pre-process to fulfill the system preconditions before being stored in the INCISIVE platform (see Section 3.1 data integration and 3.2 annotation and de identification).
- **The federated node** is the node that is hosted by the data provider where the data is stored. The data do not leave the related premises; therefore, the data partners keep full control.
- **The federated space** is the Cloud environment that contains the centralized services required to offer the federated INCISIVE functionalities.

Parts 2 and 3 constitute the repository itself.

The architecture is subject to changes and updates in order to integrate all relevant user requirements and specifications. A revision is currently underway towards a hybrid architecture that will also include a central data repository node in order to allow for more possibilities in data reusability and flexibility in data exploitation. This approach will allow very large data collections to be stored locally at sites, address data privacy and sovereignty issues, and make use of real-time Cloud high-performance resources for specific uses and specific datasets.

To date, the INCISIVE federated repository is composed of multiple interconnected nodes; hence, while physical entities of the data providers/consumers of the INCISIVE consortium are distributed around Europe, they can be virtually connected through INCISIVE interfaces, allowing the seamless and authorized access to secure data. Each data provider constitutes one node of the decentralized system, and all nodes are equally responsible for contributing to the INCISIVE architecture's overall goal, e.g., health data sharing and management for cancer business intelligence purposes. Data sources are controlled locally by their owners and regulations, and policy restrictions including GDPR are enforced appropriately without the data leaving the local premises, which in many cases is strongly preferred by hospitals and data owners in general.

Multiple databases of cancer data/images operate as a single virtual one and provide a unified source of data for front-end applications and services, e.g., AI development for cancer management purposes. The only prerequisite is that all data from the local nodes is converted to the INCISIVE CDM (in Section 3.3).

When authorized, INCISIVE users can perform queries on the data within the data-federated repository thanks to specific search modalities, such as the gender, age range, cancer type, cancer stage, months of observation, image modality, genomic data, treatment/therapy, datasets with full cases, data providers/research group, and dataset country of origin.

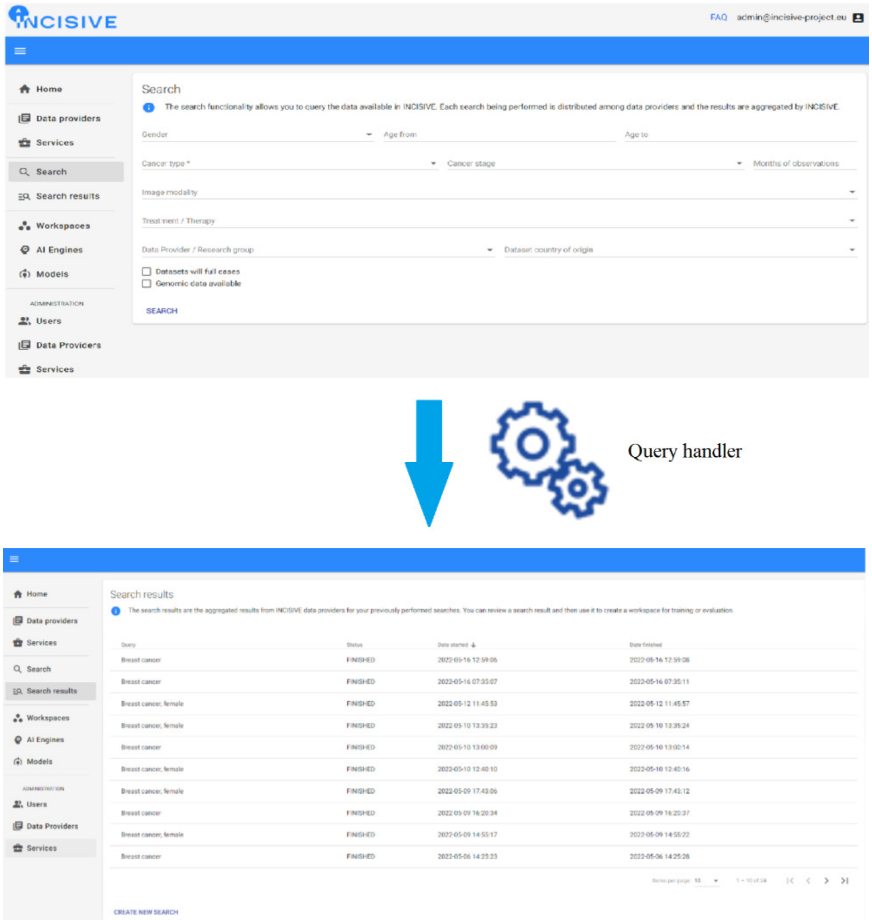
The results retrieved from each database source in the federation are then aggregated and returned to the users who submitted the query. The data never leave the physical entities that hold it. Instead, the data are "visited" and only the computed answers to the query are brought back via the INCISVE search engine API to the user (see Figure 8). If the need arises, it is also possible for the data consumers to access and examine the data stored in each federated node, provided that all relevant legal and ethical issues are previously addressed.

At the time of writing this article, the INCISIVE repository includes data from 7391 patients and 13,976 imaging examinations, covering four cancer types e.g., breast, lung, prostate, and colorectal cancer. The data come from nine data providers and five different countries, with the goal of reaching an amount of 20,000 imaging examinations over the next 2 years and much more in the long term.

The first phase of the data collection in INCISIVE was the retrospective data collection. All nine partners with the role of providing data in the project collected and prepared their data based on the data preparation procedure described in the previous session. The current status of the data stored in the infrastructure per partner is described in Table 2. The table depicts the numbers in terms of patients provided, studies provided (study is defined as an imaging examination), annotated studies provided, and the actual DICOM images. The repository will constantly be populated with more data until the end of the project.

Table 2. Current data storage status.

Partner	No of Patients	No of Studies	No of Annotated Studies	No of Images
AUTH	25	85	71	36,345
DISBA	11	26	19	14,886
GOC	67	297	-	43,391
HCS	1950	2326	1309	10,326
UNITOV	11	615	-	30,331
UNS	1392	4704	666	908,580



Search

The search functionality allows you to query the data available in INCISIVE. Each search being performed is distributed among data providers and the results are aggregated by INCISIVE.

Gender
Age from
Age to
Cancer type *
Cancer stage
Months of observations
Image modality
Treatment / Therapy
Data Provider / Research group
Dataset country of origin

□ Datasets will full cases
□ Genomic data available

SEARCH

Search results

The search results are the aggregated results from INCISIVE data providers for your previously performed searches. You can review a search result and then use it to create a workspace for training or evaluation.

Query	Status	Date started ↓	Date finished
Breast cancer	FINISHED	2022-05-16 12:59:06	2022-05-16 12:59:08
Breast cancer	FINISHED	2022-05-16 07:35:07	2022-05-16 07:35:11
Breast cancer, female	FINISHED	2022-05-12 11:45:53	2022-05-12 11:45:57
Breast cancer, female	FINISHED	2022-05-10 13:35:23	2022-05-10 13:35:24
Breast cancer	FINISHED	2022-05-10 13:00:09	2022-05-10 13:00:14
Breast cancer, female	FINISHED	2022-05-10 12:40:10	2022-05-10 12:40:16
Breast cancer, female	FINISHED	2022-05-09 17:43:06	2022-05-09 17:43:12
Breast cancer	FINISHED	2022-05-09 16:20:34	2022-05-09 16:20:37
Breast cancer, female	FINISHED	2022-05-09 14:55:17	2022-05-09 14:55:22
Breast cancer	FINISHED	2022-05-06 14:25:23	2022-05-06 14:25:28

Items per page: 10 1 – 10 of 24

CREATE NEW SEARCH

Figure 8. INCISIVE UI search process (data search and search results pages).

4.3. INCISIVE Retrospective Data—Breast Cancer Mammograpy Collection

The INCISIVE retrospective data repository, as a result of multisite efforts, had to be validated for image and clinical data quality and consistency by following the introduced conventions in terms of data structure, de-identification, and annotation. The process of quality evaluation refers to all segments of the data collection process, and has to ensure the smooth and privacy-preserved usage of medical images in AI model training, as well as the end efficient use of INCISIVE AI models on new/donated images.

Mammography, as a main tool in the breast cancer screening and follow-up process, is a cheap and commonly used modality, and has been a relevant focus for AI development. In the retrospective INCISIVE breast cancer database, MG images are present for almost every breast cancer patient (Table 3) and, sometimes, where appropriate/available are followed by breast tomosynthesis (DBT), refined 3D mammography imaging. Table 3 indicates the number of MG/DBT studies (examinations), the number of 2D images, the number of images with an accompanying annotation file, and the number of images without expert annotations.

Table 3. The number of studies, images, and annotated images for MG and DBT in the INCISIVE retrospective data repository.

DP	MG Studies	MG Images	MG Annotated Images	DBT Studies	DBT Images	DBT Annotated Images
AUTH	41	160	109	0	0	0
GOC	49	190	0	4	599	0
HCS	1950	7282	1214	3	5	0
UNS	4119	15,278	655	1618	131,653	26

All MG images in the database are scanned for duplicates, as well as for the content (MG examination or accidental surgical guidance images), dynamic range, processing status (raw or preprocessed), image size, bit depth, actual dynamic range, manufacturer, and annotation labels. A variety of properties in the collected MG images can be best depicted if the manufacturers, bits available for pixel presentation, bits per pixel stored, image size, and service–object pair (SOP) class used are compared (Table 4). Hospitals can store both raw MG images and processed MG images ("for presentation"), but most often only one option is used for scarce memory resources. This information is relevant in AI pipelines and reflected through the SOP field in the DICOM header.

Table 4. Illustration of heterogeneity of MG images in the INCISIVE retrospective database with respect to manufacturer, number of bits available and used, image size and SOP.

Manufacturer	Bits Allocated	Bits Stored	Image Size	Service–Object Pair Class
CARESTREAM Rochester, NY, USA	16	12	6000 × 4735	Computed Radiography Image Storage
FUJIFILM corporation Midtown West, Tokyo Midtown Akasaka, Minato, Tokyo, Japan	16	12, 14	5928 × 4728, 2370 × 1770, 4740 × 3540, 2964 × 2364	Breast Tomosynthesis Image Storage, Computed Radiography Image Storage, Digital Mammography X-ray Image Storage—For Presentation
HOLOGIC Inc. Marlboough, MA, USA	16	10,12	2457 × 1892, 425 × 268, 4096 × 3328, 2457 × 1996, 3328 × 2560, 2457 × 1890	Secondary Capture Image Storage, Breast Tomosynthesis Image Storage, Digital Mammography X-ray Image Storage—For Presentation
IMS GIOTTO S.P.A. Sasso Marconi (BO), Italy	16	14	2149 × 1198, 3580 × 2663, 2295 × 1120, 3580 × 2603, 1691 × 896, 3580 × 2717, 3580 × 2597, 1794 × 826, 3580 × 2812, 1817 × 876, 3580 × 2730, 3580 × 2591, 3580 × 2585, 3580 × 2657, 3580 × 2687, 1864 × 998, 3580 × 2669, 2164 × 1374, 3580 × 2675, 2287 × 1231	Breast Tomosynthesis Image Storage, Digital Mammography X-ray Image Storage—For Presentation
IMS S.R.L.	16	14	3584 × 2784, 3584 × 2736, 3584 × 2720, 3584 × 2704, 3580 × 2812, 3584 × 2768, 3584 × 2752, 3584 × 2816, 3584 × 2800	Digital Mammography X-ray Image Storage—For Presentation
LORAD, Hologic company Bedford, MA, USA	16	16	512 × 512, 1024 × 1024	Digital Mammography X-ray Image Storage—For Presentation
PHILIPS digital mammography Sweden, AB Solna, Stockholms Lan, Sweden	16	16	5355 × 4915	Digital Mammography X-ray Image Storage—For Presentation
SIEMENS Munich, Germany	16	12	1882 × 1325, 3164 × 2364, 3241 × 2203, 2577 × 1006, 3518 × 2800	Secondary Capture Image Storage, Digital Mammography X-ray Image Storage—For Presentation
SIEMENS Healthineers Erlangen, Germany	16	12, 14	2850 × 2394, 3223 × 2842, 3798 × 3328, 3359 × 2776, 2786 × 2027, 4017 × 3328, 4083 × 3328, 4092 × 3328, 3986 × 3328, 2298 × 2118, 3647 × 3328, 3867 × 2195, 3842 × 3328, 4096 × 3328, 3328 × 2560, 2665 × 2394	Secondary Capture Image Storage

The annotation labels guide the possible AI segmentation/classification tasks for MG images; thus, the relevant database statistics refer to the number of MG images containing expert segmentation of benign, malignant, or suspicious lesions, calcifications, surgical clips, and axial lymph nodes (Table 5). However, it should be noted that besides those expert annotation files, all images in the database are accompanied with clinical data offering additional classification tasks options, such as the BIRADS, cancer type, cancer grade, molecular subtype, treatment information, time with respect to diagnoses, etc.

Table 5. Number of MG images per annotation label.

	Annotation Label						
Provider	**Suspicious/Indeterminate**	**Malign**	**Calcification**	**Surgical Clip**	**Axial Lymph Node**	**Benign**	**No Findings**
AUTH	5	28	85	28	15	15	12
HCS	160	299	558	50	249	300	7
UNS	213	53	87	86	87	52	216

The data schema used in the prospective data collection assumed that patient journey can be followed in time, and each clinical episode from diagnoses to treatment and follow-up is summarized in a certain TP defined by months from diagnoses. The information on number of TPs for each patient is relevant in order to evaluate the potential of the database for tasks related to the prediction of disease outcome, metastasis risk, treatment effects, etc. With respect to MG images, the number of TPs available for breast cancer patients is presented in Figure 9. Similarly, in Figure 10, the same information is presented for DBT images.

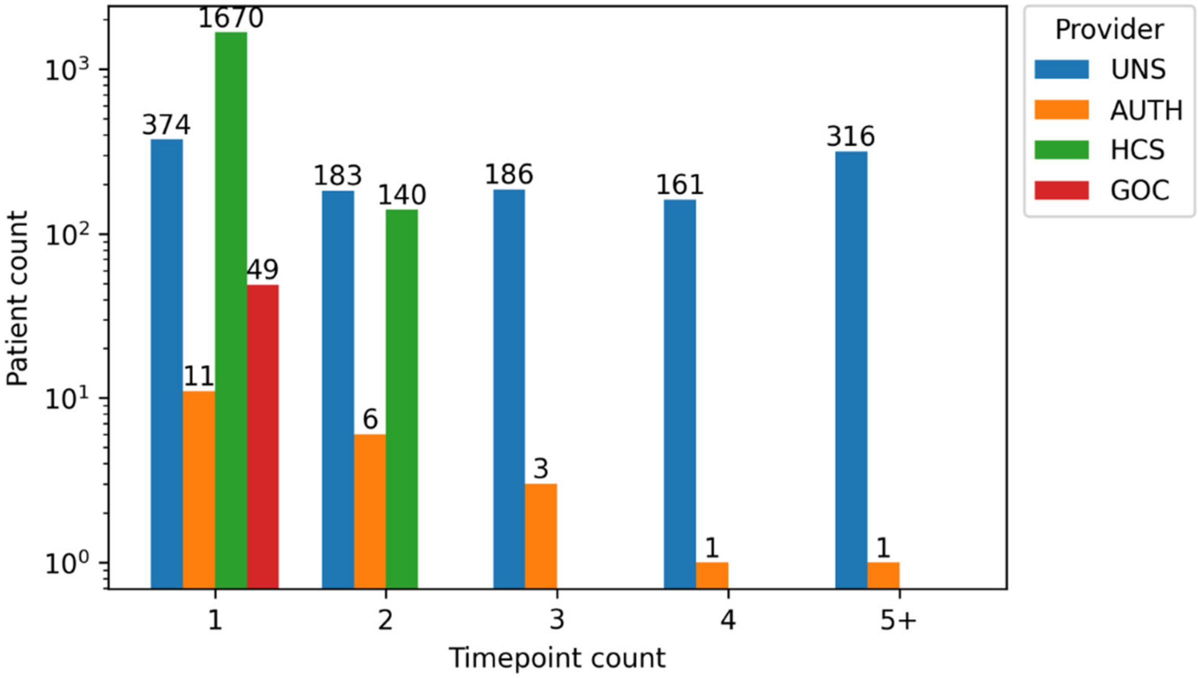


Figure 9. The number of patients with MG images in different numbers of TPs in the INCISIVE retrospective database.

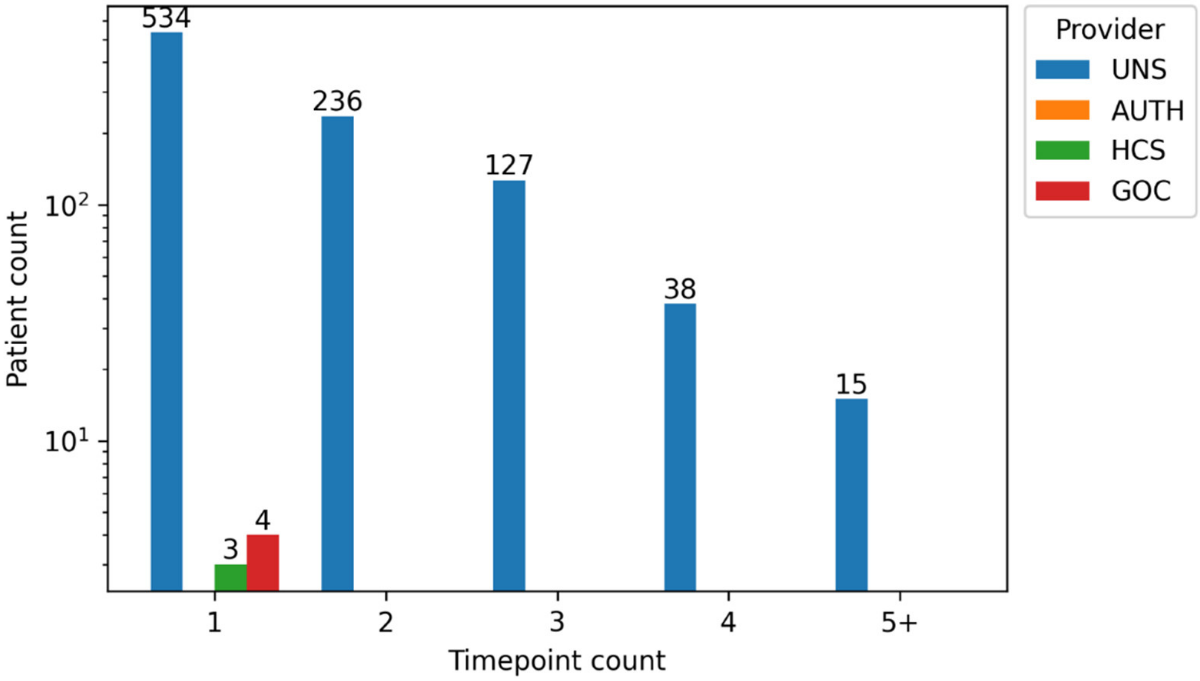


Figure 10. The number of patients with DBT images in different numbers of TPs in the INCISIVE retrospective database.

5. Description of AI Toolbox/Breast Cancer Tools

5.1. Federated Learning—Data Exploitation

As privacy is a key factor in the healthcare environment, most of the time, hospitals and other entities that store medical data face difficulties or even do not want to share their data with researchers. For this reason, research on fields such as breast cancer imaging analysis can be hard to perform, even though there are an increasing number of open datasets. However, federated learning [47] offers a way of training machine learning models without the need of sharing data.

5.1.1. How It Is Performed

Federated learning is a kind of distributed machine learning technique which involves distributing the training process where the data is produced or stored. In this sense, the machines where the data are stored are used to train the required models. By doing so, the data are always kept at the institution premises and not shared externally, but can be used to build models. Inside INCISIVE, there are several components that manage the lifecycle of the learning process:

- Orchestrator: in charge of receiving the training requests and deploying the required components in the central node and in the federated nodes.
- Model-as-a-service (MaaS): in charge of storing models and performing inference.
- AI federator: in charge of managing the federated learning process (send training requests, receive weights, and merge models).
- AI engine: in charge of performing the actual training at each federated node. This component is composed of auxiliary components to provide data and retrieve results from a standard machine learning application agnostic of the federated learning and the infrastructure.

The process, illustrated in Figure 11, is as follows.

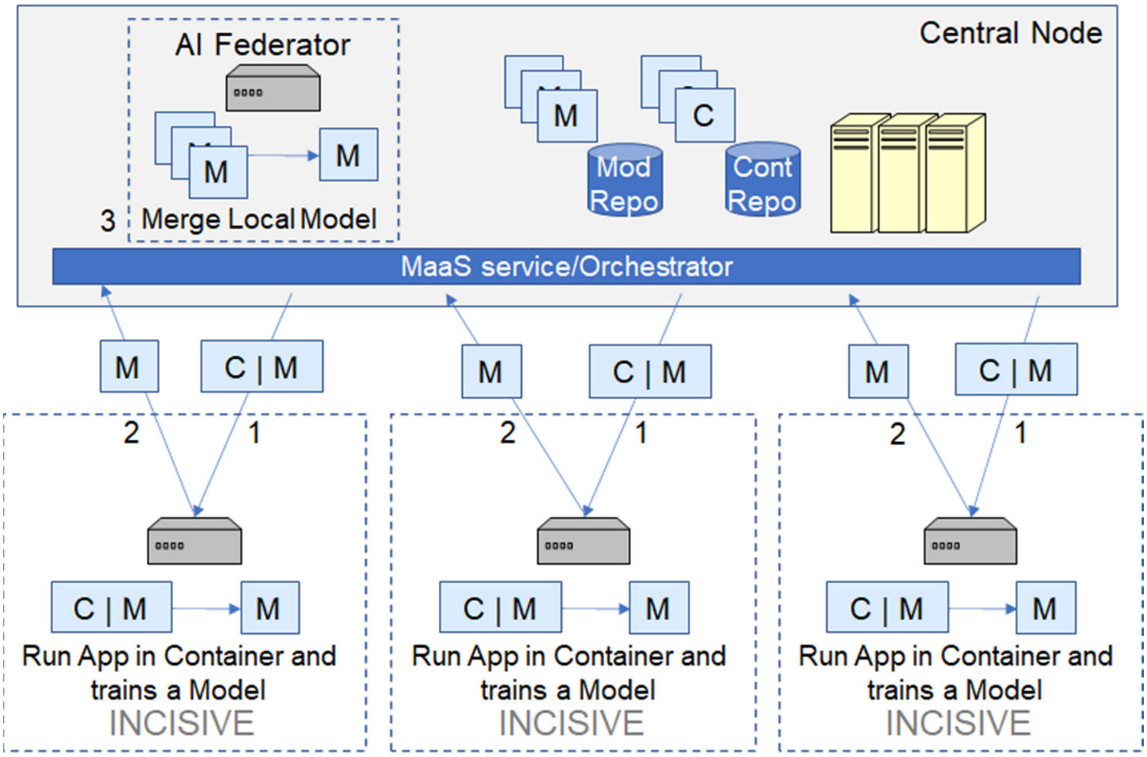


Figure 11. Federated sharing overall process. C stands for containers (required containerized application, e.g., a docker image) and M for models. Notice that the container is only required to be sent once at each federated node. The code hosted in the container is run using the input model and the data is stored in the node, producing a new model as the output.

1. A train request is received in the UI.
2. The orchestrator analyzes the request, and a specific AI federator and AI engines of the federated nodes are deployed.
3. The AI federator initializes the model or loads a previously trained model from the MaaS, depending on if training from scratch or training from a pretrained model is requested.
4. The AI engine receives the model and trains it, using local data available.
5. The AI federator receives the trained models and merges them using a merging scheme, e.g., FedAvg [47].
6. In case another federated learning round is required, the process restarts from Step 4, sending the new merged model to the AI engines.
7. Finally, the model is stored in the MaaS.

The iterative process of the federated learning rounds is required to converge in a solution that is satisfying for all the different data providers that participate in the federated learning process. Notice that if an excessive number of epochs is used in every federated learning training round, the federated nodes can produce models overfitted to the local data. This could mean that the merged model does not converge into a good and general solution.

The work of Linardos et al. [48] on cardiovascular diseases shows that this kind of approach is well fitted for training deep learning models for image analysis. In particular, this work shows that the accuracy of the federated learning trained model is close to training in a single computer with all the data. In particular, in some cases, the results are better than the traditional training approach. This could be due to the regularizing effect of merging models studied in works such as Izmailov et al. [49].

5.1.2. Hybrid Infrastructure and Federated Learning

Federated learning enables training without having the data directly accessible. This is positive for data privacy; however, it works in detriment of ease of use. For this reason, in the INCISIVE project, a hybrid infrastructure is planned so that data can be shared in a central repository. In particular, this is offered from the start in a Cloud-based infrastructure (temporary infrastructure) so that the AI developers from the project can start their work as soon as possible. This way, the developers and researchers can test their code and check the cases in which their models fail.

With this hybrid infrastructure, INCISIVE will be able to also offer storage for those data providers that do not have an infrastructure to hold the federated learning components. This makes the infrastructure accessible and versatile.

5.2. Preliminary Results—Initial Analysis on Mammography

Medical images differ to natural images in many aspects, and most importantly by their high resolution and very small regions of interest (ROIs). For these reasons, if deep learning models are trained and developed for natural images [50–53], they assume a severe down-sampling of medical images for memory constraints and offer too coarse localization [54]. In high-resolution MG images, in early asymptomatic stages relevant for breast cancer screening, ROIs are usually very small, with subtle differences to the normal tissue and often sparsely distributed [54]. In order to deliver an accurate decision, an MG classifier should not only detect suspicious changes (e.g., lesions and calcifications), but also take into consideration their characteristics (e.g., intensity, shape, and position) and some global features (e.g., breast fibro-glandular tissue density and pattern) [55,56].

Based on the extensive literature review, in the INCISIVE, the focus will be placed on few recently proposed architectures for the segmentation of lesions and malignancy classification which will be trained or evaluated on both internally collected data and available MG open datasets. The relevant line of research is based on the NYU breast cancer screening dataset (NYUBCS) [57] comprising 1 million MG images, which facilitated the development of the classification models being fully trained on MG images. In the INCISIVE, several explainable classification models developed on NYUBSC will be tested. The breast-label classification model, a globally aware multiple instance classifier (GMIC) [58], will serve as a baseline, as it achieves AUC = 0.93 on NYUBSC. As the segmentation of lesions is one of the main tasks in INCISIVE, the concept of weakly supervised localization (WSL) [59] has been explored in order to alleviate the problem of expensive and time-consuming experts' pixel-level annotation. Aligned with WSL, the global–local activation map (GLAM) model [54] uses only a breast-level label to produce a fine segmentation at a 300-times-higher resolution of saliency maps when compared with GMIC [58], and an improvement of 20%, as measured by the DICE index on the NYUBCS MG database. So far, both GMIC and GLAM input pipelines were formed with the specific processing steps described in [54,58], respectively. Additional steps were made to correct any potential differences of INCISIVE MG images from the training NYUBSC data in terms of contrast and dynamic ranges, while following other noted required attributes from the NYUBCS image selection process. By thresholding the corresponding output saliency maps, specific regions can be segmented and identified as benign or malignant, as determined by the model.

Besides these models trained on the closed NYUBCS MG database, and clearly focused on MG classification and WSL, there are some other more general approaches to medical image segmentation, not restrained to MG. In INCISIVE, the medical transformers (MTs) [60] will be trained and evaluated using the INCISIVE data on images with available pixel-level annotation. The pipelines have already been prepared and efforts directed to the harmonization of image attributes from different data providers within INCISIVE. The preprocessed MG images are automatically cropped to ROIs and rescaled to a modest 256 × 256 size required at input.

nnU-Net is a recently proposed medical image segmentation method which aims to alleviate the need for a highly specialized solution for numerous distinctive problems in medical imaging [61]. It should be automatically configurable for any new task, adjusting all relevant steps: preprocessing, network architecture, training, and post-processing. nnU-net code has been modified to work on single-view MG images involving the conversion of 2D DICOM images into 3D NIFTI images and their organization into an appropriate folder structure. Images were normalized, cropped to ROIs, and rescaled to the same size (320 × 416). Initially, the training was attempted on a smaller dataset in order to evaluate available implementation [61]. Currently, the code is being decomposed and adjusted to facilitate for the planned federated data storage.

5.2.1. Data Quality and Related Pre-Processing

INCISIVE bases its AI development on both open datasets and collected data. In order to achieve good generalization of AI models, certain robustness has to be ensured with respect to input image quality. For these reasons, the heterogeneity of collected MG images in all relevant parameters has to be investigated. The guidelines for quality assurance in breast cancer screening [62] set the clear minimum rules on how MG imaging is to be performed, covering all physical aspects of the imaging process, with the calibration, radiation dose, and exposure time being the main determinants of the MG image quality [63]. Once an MG image is produced, the image quality can be improved by relying on digital image processing tools, focusing on contrast enhancement and noise removal [62].

When creating the INCISIVE database, MG images could be uploaded as raw images, i.e., produced by the detector, corrected for gain and offset, and processed where contrast enhancement and noise removal have already been applied by manufacturer software. In the data collection process, image quality has not been automatically checked, but the correspondence between clinical metadata and imaging data tested in each time point. For these reasons, all MG images in the repository had to be automatically assessed for quality to ensure their usability in AI models and to set some additional rules for the data donorship schema.

Guided by the experiences in collecting the largest MG image database so far (NYUBCS) [57], the DICOM header of MG images had to be inspected to remove all images that did not relate to the MLO or CC view of the left or right breast. It should be noted that different types of errors during the examination are possible, yet these images are not usually deleted by the hospital technicians and can be accidentally uploaded. When filtering images, there are multiple possible reasons to discard an image, such as poor image resolution, views other than CC or MLO, a lack of important information in the DICOM header in relation to the SOP, image type, if an image is scanned for surgical guidance, and if the proportion of nonzero pixels is more than 95% or less than 5%.

In order to provide uniform image appearance at the model input, multiple contrast enhancement pre-processing methods have to be evaluated jointly with AI model training in order to select the most appropriate procedure for image pre-processing. For every image, information on the basic image attributes will be derived from its DICOM header (or alternatively automatically if certain DICOM fields are missing), and preprocessing will be dependent on the image attributes. Currently, in the INCISIVE database, all MG images uploaded are denoted as processed (for presentation); however, despite this field, they do not have uniform appearance and might even be presented as negative. Figure 12 provides some examples of INCISIVE MG images from different data providers. Figure 13a shows an MG image with an automatically identified breast region (ROI), while panel (b) presents the image histogram (in blue), and a proposed piecewise linear intensity transformation in red. The enhanced ROI is presented in Figure 13d with a histogram indicating an improved utilization of the available intensity range (Figure 13c). An automated procedure estimating the dynamic range of an image, as well as the contrast and noise levels, will determine the transformation needed to harmonize the image appearance.

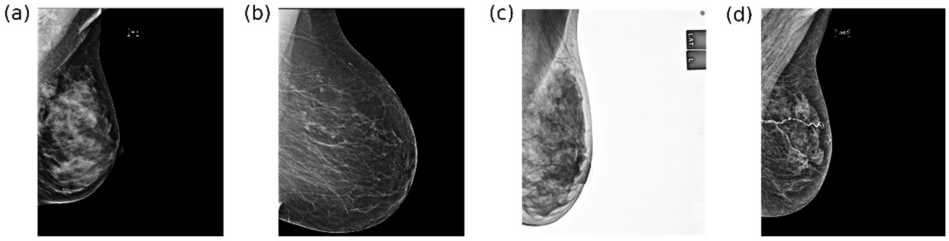


Figure 12. MG images from the INCISIVE retrospective database in original size, all intended for presentation, and originating from different data providers: (**a**) AUTH, (**b**) GOC, (**c**) HCS, and (**d**) UNS. Note that images from different providers and different manufacturers might undergo different processing steps from a raw image format, which is an unavailable piece of information.

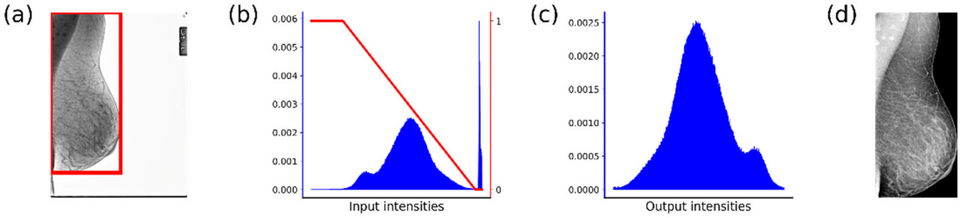


Figure 13. (**a**) An MG image in original size with an indicated region of interest, (**b**) a histogram of the original image (in blue) and intensity transformation applied for image enhancement (in red), (**c**) a histogram after pre-processing, and (**d**) a preprocessed image.

5.2.2. Annotation Style

The medical professionals from the DP institutions annotated a fraction of the uploaded MGs using a predefined annotation protocol and contour segmentation. As presented in Table 1, the lesions/regions in the breast could be marked as: benign, malignant, or suspicious/indeterminate. Based on the analysis of the manually segmented MG images, the proportion of malignant and suspicious labels varies between data providers (Table 4). This might stem from the fact that certain clinicians evaluated images blindly, unaware of the accompanying clinical data, in order to make a fair and comparable decision on those made by image-based AI models, while others might make annotations with awareness of the clinical background and biopsy result. This problem might be alleviated by information extraction from clinical metadata containing the biopsy label for each patient in case of the presence of malignant/suspected or benign lesions in the accompanying images. The biopsy results will be considered as ground truth in image segmentation and classification tasks.

Another interesting research avenue is related to the different style of manual segmentation, as experts can be more or less precise when contouring the lesions, which directly influences all quantitative measures for automatic image segmentation, such as the DICE coefficient. In order to approach this problem, the usability and efficacy of unsupervised solutions to infer the ground truth and introduce the segmentation quality control will be explored [64]. Moreover, the harmonization of the annotation procedure and alignment of multiple experts will be additionally achieved through a more detailed consensus in MG image annotation, which will serve as a guideline in a data donorship schema.

6. Conclusions and Next Steps

In this manuscript, the INCISIVE platform was introduced, which will be utilized for the development and evaluation of AI and data analytics in order to recognize complex patterns in images, such as mammography, while also providing the ability to share medical health data in order to address data availability challenges. For the above reasons, the data should be collected in such a way that all legal, privacy, and ethical challenges are taken into consideration. The approach to these challenges via the adopted legal framework in INCISIVE has been elaborated and explained. In addition, due to the divergence of the data types and data sources, the article defines the process of data collection, data preparation, and data integration within INCISIVE, which is monitored using the developed tools for data quality assessment. An automated quality control should identify data, without curation and integration guidelines.

For the breast cancer case, the article comprehensively presents the data/image types to be integrated, data collection steps, the guidelines for annotation of mammography examinations, and the process of uploading de-identified data. The INCISIVE retrospective collection of MG images is described and illustrated from several perspectives for possible AI use cases.

The vision of the INCISIVE pan European federated repository of health images is outlined, and the underlying platform is presented, along with the adopted principles and solutions for federated data sharing. We demonstrate the potential for developing and evaluating the AI toolbox, thus explaining the process and basic components used to facilitate federated learning, besides the basic centralized learning approach.

This study sets the stage for the training and evaluation of AI models based on a set of homogenized and curated datasets; however, at the current project phase, it does not demonstrate AI model performance results, nor does it support the evidence for comparison of a centralized vs. federated approach in model development. Despite the data collection guidelines envisaged following the patients in multiple time points, the largest number of patients collected in the retrospective study has data which are only available in baseline, or alternatively in one more follow-up TP. This limitation will influence the development of prognostic models, and predictions relating to disease progression and treatment efficiency over time. The prospective study protocols overcome this limitation and should offer complete information of the patient's journey.

Moreover, this paper does not discuss the consortium efforts invested and the methodology used in the analysis of user needs, as well as the motivation and criteria associated with the user acceptance of AI-driven solutions, which guided the platform development and service design. It is worth mentioning that the adopted data harmonization, technological approaches, and platform architecture are a result of the collaborative efforts and experience of the project partners; thus, they do not illustrate a unique approach to tackle the challenges of Big Data integration for AI use in healthcare.

Overall, by increasing data availability and reducing economic barriers, INCISIVE should enhance the democratization of narrow AI through the integration of existing datasets, and provide inexpensive local access to high-performance hardware in federated and central nodes. The accessibility and reusability are targeted through: (1) data availability—the development of the EU-wide repository using a federated storage approach that allows data owners to reuse their data and integrate them with other datasets; (2) data shareability—the INCISIVE donorship scheme tackles interoperability issues and makes sharing possible, block-chain safe, and regulation aligned; and (3) affordability—reduced experimentation costs with AI as a service solution to lower human resource costs for image annotation, de-identification, and reconstruction.

The initial collection of retrospective data will be amended by additional datasets required for the validation and corrections of developed AI models. The INCISIVE AI toolbox should be made available for clinical validation, which includes the efficiency, consistency, and generalizability of results. Even so, there is a high demand and increased interest for AI/ ML applications in the clinical setting. The INCISIVE project aims to strengthen the results and improve the acceptance of AI by providing explanations for such results, encouraging domain experts to accept and use AI for cancer diagnosis. Moreover, global organizations must accept AI in clinical validation studies and be prepared to improve guidance and regulation surrounding AI in oncology, along with formal integrated training for medical purposes.

Author Contributions: Conceptualization, A.G.-T., D.D., S.M.-A., S.A., T.L.-T., A.R., J.L.B., C.B., N.M., P.B., I.C., D.F., I.H., M.K.-C., A.K., A.L., M.L., S.N.-G., Y.R., E.S., O.T., K.V. and B.A.; methodology, F.A., D.D., S.M.-A., S.A., A.G.-T., I.L., N.J., T.L.-T., C.B., N.M., P.B., I.C., D.F., I.H., A.K., A.L., Y.R., E.S., O.T., K.V. and B.A.; software, A.G.-T., F.A., C.B., N.M., P.B., A.K., E.S. and B.A.; formal analysis, I.L., N.J. and T.L.-T.; investigation, A.G.-T., C.B., N.M., P.B., D.F., I.H., R.K., S.N.-G., Y.R., and B.A.; data curation, D.D., I.L., and N.J.; writing—original draft preparation, C.S., O.B., A.C., T.L.-T., A.G.-T., C.B., N.M., P.B., I.C., I.H., R.K., A.K., Y.R., E.S., O.T., K.V. and B.A.; writing—review and editing, C.S., O.B., A.G.-T., C.B., N.M., P.B., I.C., I.H., A.K., Y.R., E.S., O.T. and B.A.; visualization, I.L., A.K. and T.L.-T.; supervision, J.L.B. and A.C.; funding acquisition, J.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding mainly from the European Union's Horizon 2020 research and innovation program under grant agreement no 952179. It was also partially funded by the Ministry of Economy, Industry, and Competitiveness of Spain under contracts PID2019-107255GB and 2017-SGR-1414.

Acknowledgments: The authors would also like to acknowledge Kostas Xythalis, Irine Georgiou, and Melpomeni Kosmidou for their provision of mammography data. In addition, the authors would also like to acknowledge the work of the rest of the INCISIVE partners.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bi, W.L.; Hosny, A.; Schabath, M.B.; Giger, M.L.; Birkbak, N.J.; Mehrtash, A.; Allison, T.; Arnaout, O.; Abbosh, C.; Dunn, I.F.; et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J. Clin.* **2019**, *69*, 127–157. [CrossRef] [PubMed]
2. Lima, Z.S.; Ebadi, M.R.; Amjad, G.; Younesi, L. Application of Imaging in Breast Cancer Detection: A Review Article. *Maced. J. Med. Sci.* **2019**, *7*, 838–848. [CrossRef] [PubMed]
3. Munir, K.; Elahi, H.; Ayub, A.; Frezza, F.; Rizzi, A. Cancer diagnosis using deep learning: A bibliographic review. *Cancers* **2019**, *11*, 1235. [CrossRef] [PubMed]
4. Michael, E.; Ma, H.; Li, H.; Kulwa, F.; Li, J. Breast cancer segmentation methods: Current status and future potentials. *BioMed Res. Int.* **2021**, *2021*, 9962109. [CrossRef]
5. Immonen, E.; Wong, J.; Nieminen, M.; Kekkonen, L.; Roine, S.; Törnroos, S.; Lanca, L.; Guan, F.; Metsälä, E. The use of deep learning towards dose optimization in low-dose computed tomography: A scoping review. *Radiography* **2021**, *28*, 208–214. [CrossRef]
6. Davenport, T.; Kalakota, R. The potential for artificial intelligence in healthcare. *Future Health J.* **2019**, *6*, 94–98. [CrossRef]
7. Mahadevaiah, G.; Rv, P.; Bermejo, I.; Jaffray, D.; Dekker, A.; Wee, L. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. *Med. Phys.* **2020**, *47*, e228–e235. [CrossRef]
8. Iqbal, M.J.; Javed, Z.; Sadia, H.; Qureshi, I.A.; Irshad, A.; Ahmed, R.; Malik, K.; Raza, S.; Abbas, A.; Pezzani, R.; et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. *Cancer Cell Int.* **2021**, *21*, 270. [CrossRef]
9. Tarumi, S.; Takeuchi, W.; Chalkidis, G.; Rodriguez-Loya, S.; Kuwata, J.; Flynn, M.; Kawamoto, K. Leveraging artificial intelligence to improve chronic disease care: Methods and application to pharmacotherapy decision support for type-2 diabetes mellitus. *Methods Inf. Med.* **2021**, *60*, e32–e43. [CrossRef]
10. Miller, D.D.; Brown, E.W. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am. J. Med.* **2018**, *131*, 129–133. [CrossRef]
11. Meskó, B.; Görög, M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit. Med.* **2020**, *3*, 126. [CrossRef]
12. Bohr, A.; Memarzadeh, K. The rise of artificial intelligence in healthcare applications. *Artif. Intell. Healthc.* **2020**, 25–60. [CrossRef]
13. Langlotz, C.P.; Allen, B.; Erickson, B.J.; Kalpathy-Cramer, J.; Bigelow, K.; Cook, T.S.; Flanders, A.E.; Lungren, M.P.; Mendelson, D.S.; Rudie, J.D.; et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology* **2019**, *291*, 781–791. [CrossRef] [PubMed]
14. Gerke, S.; Minssen, T.; Cohen, G. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artif. Intell. Healthc.* **2020**, 295–336. [CrossRef]
15. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation, GDPR). Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 18 August 2022).

16. Geis, J.R.; Brady, A.P.; Wu, C.C.; Spencer, J.; Ranschaert, E.; Jaremko, J.L.; Langer, S.G.; Kitts, A.B.; Birch, J.; Shields, W.F.; et al. Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *J. Am. Coll. Radiol.* **2019**, *16*, 1516–1521. [CrossRef]
17. Van Veen, E.; Boeckhout, M.; Schlünder, I.; Boiten, J.W.; Dias, V. Joint controllers in large research consortia: A funnel model to distinguish controllers in the sense of the GDPR from other partners in the consortium [version 1; peer review: Awaiting peer review]. *Open Res. Eur.* **2022**, *2*, 80. [CrossRef]
18. Scheibner, J.; Raisaro, J.L.; Troncoso-Pastoriza, J.R.; Ienca, M.; Fellay, J.; Vayena, E.; Hubaux, J.P. Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis. *J. Med. Internet Res.* **2021**, *23*, e25120. [CrossRef]
19. Wan, Z.; Hazel, J.W.; Clayton, E.W.; Vorobeychik, Y.; Kantarcioglu, M.; Malin, B.A. Sociotechnical safeguards for genomic data privacy. *Nat. Rev. Genet.* **2022**, *23*, 429–445. [CrossRef]
20. Thorogood, A.; Rehm, H.L.; Goodhand, P.; Page, A.J.H.; Joly, Y.; Baudis, M.; Rambla, J.; Navarro, A.; Nyronen, T.H.; Linden, M.; et al. International federation of genomic medicine databases using GA4GH standards. *Cell Genom.* **2021**, *1*, 100032. [CrossRef]
21. 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. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
22. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act). 2022. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868&qid=1660831508274 (accessed on 18 August 2022).
23. Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act). Available online: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A842%3AFIN (accessed on 18 August 2022).
24. Proposal for a Regulation of the European Parliament and of the Council laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act, AI Act) and Amending Certain Union Legislative Acts, COM/2021/206 Final. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (accessed on 18 August 2022).
25. Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space, COM/2022/197 Final. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197 (accessed on 18 August 2022).
26. Kristensen, F.B.; Husereau, D.; Huić, M.; Drummond, M.; Berger, M.L.; Bond, K.; Augustovski, F.; Booth, A.; Bridges, J.F.P.; Grimshaw, J.; et al. Identifying the Need for Good Practices in Health Technology Assessment: Summary of the ISPOR HTA Council Working Group Report on Good Practices in HTA. *Value Health* **2019**, *22*, 13–20. [CrossRef] [PubMed]
27. Kosvyra, A.; Filos, D.; Fotopoulos, D.; Olga, T.; Chouvarda, I. Towards Data Integration for AI in Cancer Research. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico City, Mexico, 1–5 November 2021; IEEE: New York, NY, USA, 2021.
28. ITK-Snap Version 3.8.0. Available online: http://www.itksnap.org/pmwiki/pmwiki.php?n=Downloads.SNAP3 (accessed on 24 August 2022).
29. Kosvyra, A.; Filos, D.; Fotopoulos, D.; Olga, T.; Chouvarda, I. Data Quality Check in Cancer Imaging Research: Deploying and Evaluating the DIQCT Tool. In Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, UK, 11–15 July 2022.
30. DICOM Security and System Management Profiles. Available online: https://dicom.nema.org/medical/dicom/current/output/chtml/part15/chapter_e.html (accessed on 24 August 2022).
31. MIRC CTP-MircWiki. Available online: https://mircwiki.rsna.org/index.php?title=MIRC_CTP (accessed on 24 August 2022).
32. Graham, R.N.; Perriss, R.W.; Scarsbrook, A.F. DICOM demystified: A review of digital file formats and their use in radiological practice. *Clin. Radiol.* **2005**, *60*, 1133–1140. [CrossRef] [PubMed]
33. DICOM Standard PS 3.15 Digital Imaging and Communications in Medicine (DICOM), Part 15: Security and System Management Profiles. Available online: http://dicom.nema.org/dicom/2013/output/chtml/part15/PS3.15.html (accessed on 24 August 2022).
34. McDonald, C.J.; Huff, S.M.; Suico, J.G.; Hill, G.; Leavelle, D.; Aller, R.; Forrey, A.; Mercer, K.; DeMoor, G.; Hook, J.; et al. LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clin. Chem.* **2003**, *49*, 624–633. [CrossRef] [PubMed]
35. SNOMED. Available online: https://www.snomed.org/ (accessed on 24 August 2022).
36. Esthesis Platform. Available online: https://www.eurodyn.com/rnd-product/esthesis/ (accessed on 24 August 2022).
37. OMOP Common Data Model. Available online: https://www.ohdsi.org/data-standardization/the-common-data-model/ (accessed on 24 August 2022).
38. The CTP DICOM Anonymizer. Available online: https://mircwiki.rsna.org/index.php?title=The_CTP_DICOM_Anonymizer (accessed on 24 August 2022).
39. Yushkevich, P.A.; Piven, J.; Hazlett, H.C.; Smith, R.G.; Ho, S.; Gee, J.C.; Gerig, G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef]
40. Schünemann, H.J.; Lerda, D.; Quinn, C.; Follmann, M.; Alonso-Coello, P.; Rossi, P.G.; Lebeau, A.; Nyström, L.; Broeders, M.; Ioannidou-Mouzaka, L.; et al. Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines. *Ann. Intern. Med.* **2019**, *172*, 46–56. [CrossRef]

41. Runowicz, C.D.; Leach, C.R.; Henry, N.L.; Henry, K.S.; Mackey, H.T.; Cowens-Alvarado, R.L.; Ganz, P.A. American cancer society/American society of clinical oncology breast cancer survivorship care guideline. *CA A Cancer J. Clin.* **2016**, *66*, 43–73. [CrossRef]
42. Lazic, I.; Jakovljevic, N.; Boban, J.; Nosek, I.; Loncar-Turukalo, T. Information extraction from clinical records: An example for breast cancer. In Proceedings of the 21st IEEE Mediterranean Electrotechnical Conference, Palermo, Italy, 14–16 June 2022.
43. Symvoulidis, C.; Marinos, G.; Kiourtis, A.; Mavrogiorgou, A.; Kyriazis, D. HealthFetch: An Influence-Based, Context-Aware Prefetch Scheme in Citizen-Centered Health Storage Clouds. *Future Internet* **2022**, *14*, 112. [CrossRef]
44. Symvoulidis, C.; Kiourtis, A.; Mavrogiorgou, A.; Kyriazis, D. Healthcare Provision in the Cloud: An EHR Object Store-based Cloud Used for Emergency. *Healthinf* **2021**, *1*, 435–442.
45. NextCloud. Available online: https://nextcloud.com/ (accessed on 24 August 2022).
46. Intel Software Guard Extensions. Available online: https://www.intel.com/content/www/us/en/developer/tools/software-guard-extensions/overview.html (accessed on 24 August 2022).
47. McMahan, H.B.; Moore, E.; Ramage, D.; Arcas, B.A. Federated learning of deep networks using model averaging. *arXiv* **2016**, arXiv:1602.05629.
48. Linardos, A.; Kushibar, K.; Walsh, S.; Gkontra, P.; Lekadir, K. Federated Learning for Multi-Center Imaging Diagnostics: A Study in Cardiovascular Disease. *arXiv* **2021**, arXiv:2107.03901. [CrossRef]
49. Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; Wilson, A.G. Averaging weights leads to wider optima and better generalization. *arXiv* **2018**, arXiv:1803.05407.
50. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23 June–28 June 2014.
51. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
52. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
54. Liu, K.; Shen, Y.; Wu, N.; Chłędowski, J.; Fernandez-Granda, C.; Geras, K.J. Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis. *Proc. Mach. Learn. Res.* **2021**, *143*, 268. [PubMed]
55. Shen, Y.; Wu, N.; Phang, J.; Park, J.; Liu, K.; Tyagi, S.; Heacock, L.; Kim, S.G.; Moy, L.; Cho, K.; et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med. Image Anal.* **2021**, *68*, 101908. [CrossRef] [PubMed]
56. Wei, J.; Chan, H.-P.; Wu, Y.-T.; Zhou, C.; Helvie, M.A.; Tsodikov, A.; Hadjiiski, L.M.; Sahiner, B. Association of computerized mammographic parenchymal pat- tern measure with breast cancer risk: A pilot case-control study. *Radiology* **2011**, *260*, 42–49. [CrossRef] [PubMed]
57. Wu, N.; Phang, J.; Park, J.; Shen, Y.; Kim, S.G.; Heacock, L.; Moy, L.; Cho, K.; Geras, K.J. *The NYU Breast Cancer Screening Dataset V1.0*; New York University: New York, NY, USA, 2019.
58. Shen, Y.; Wu, N.; Phang, J.; Park, J.; Kim, G.; Moy, L.; Cho, K.; Geras, K.J. Globally-aware multiple instance classifier for breast cancer screening. In *International Workshop on Machine Learning in Medical Imaging*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 18–26.
59. Diba, A.; Sharma, V.; Pazandeh, A.; Pirsiavash, H.; Van Gool, L. Weakly supervised cascaded convolutional networks. *arXiv* **2017**, arXiv:1611.08258.
60. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 36–46.
61. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
62. Amendoeira, I.; Perry, N.; Broeders, M.; de Wolf, C.; Törnberg, S.; Holland, R.; von Karsa, L. *European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis*; European Commission: Brussels, Belgium, 2013.
63. Williams, M.B.; Raghunathan, P.; More, M.J.; Seibert, J.A.; Kwan, A.; Lo, J.; Samei, E.; Ranger, N.T.; Fajardo, L.L.; McGruder, A.; et al. Optimization of exposure parameters in full field digital mammography. *Med. Phys.* **2008**, *35*, 2414–2423. [CrossRef]
64. Audelan, B.; Delingette, H. Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model. *Med. Image Anal.* **2020**, *68*, 101895. [CrossRef]