Topic: Central repository of digital pathology slides to support the development of artificial intelligence tools

All information regarding future IMI Call topics is indicative and subject to change. Final information about future IMI Calls will be communicated after approval by the IMI Governing Board.

Topic details

Action type	Research and Innovation Action (RIA)
Submission and evaluation process	2 stages

Specific challenges to be addressed

Although pathology is the cornerstone of the workup of many diseases such as cancer, autoimmune diseases, and transplant rejection, it still relies heavily on the subjective interpretation of a histology sample by a qualified pathologist who captures observations and conclusions in a report. Once the observations are captured, the slides are archived and only the pathologist's report and diagnoses (considered as raw data in good laboratory practice (GLP) nonclinical studies) remain accessible. Therefore, significant information from the histology slides is no longer easily available. This hinders the discovery of new clinico-pathological entities that are relevant to patients' prognosis and treatment.

The recent developments of high-throughput slide scanners offer a possibility for making the entire information contained in the millions of glass slides produced every year, available for search. Ensuring storage and access to digital slides will overcome the current limitations to accessing and sharing pathology material together with the associated metadata. It will facilitate case consultation, help identify sub-types of diseases, assess the translatability of nonclinical safety observations and animal models, and thereby rationalise the design of clinical trials and the use of animal models.

The rise of deep learning and its unexpected ease at interpreting images offer unprecedented opportunities to develop tools for automated detection, classification and quantification of abnormalities in tissues. Hence, many initiatives are already looking at utilising histopathology slides in a digital format as a source of data for biomedical research. Current research focuses on a relatively reduced set of diseases and/or are fragmented and geographically limited, which may hinder their ability to deliver outside of much-targeted applications.

This is mostly because, although clinically relevant and efficient, disease-centric models cannot be easily expanded towards more general purposes.

However, the full transformative potential of deep learning applied to histopathology goes far beyond what is presently undertaken. In the future, it will provide the pathologist with smart suggestions regarding diagnoses and mechanistic or therapeutic hypotheses (predict patient's outcomes and responses to treatment), significantly improving overall patient safety and diagnosis. To achieve this ambitious goal, a much larger series of slides offering a broader coverage of tissues and lesions is required. Whereas such coverage may be difficult to achieve solely with clinical material, nonclinical toxicology studies provide an incredibly valuable and abundant source of histopathology slides, comprising all the normal tissues from multiple species, and a large diversity of lesions. As these lesions are similar to those seen in clinical practice, but in a more pure form, and at stages rarely encountered in humans, they will be a great help for the community developing artificial intelligence (AI). They will also likely offer an opportunity to expedite the development of assisted diagnosis tools applicable to nonclinical safety studies and clinical practice.

Need and opportunity for public-private collaborative research

The refinement of the pharmaco-therapeutic armamentarium requires the improvement of disease classification and of diagnostic and prognostic criteria. This is an ongoing effort in several areas of medicine.

Indicative text

However, for many diseases, it is hampered by limited access to large histopathology series and the absence of reliable quantitative methods. To overcome these obstacles, it is necessary to make large sets of histopathology slides accessible to the medico-scientific community in a digital form.

The current efforts in the field of machine learning and histopathology focus on the development of disease-specific models. Although their potential clinical utility is compelling, such models are limited to a particular tissue. The development of holistic models is necessary to support improvements in disease classification and translational research, which will in turn accelerate the discovery of new clinico-pathological entities and provide assisted diagnostics tools.

The magnitude of the challenges addressed by the Call topic is such that they cannot be addressed solely by the academic or industry sectors.

Firstly, it requires the collection of sufficiently large sets of histology data along with associated clinical information. The pharmaceutical industry will provide high-quality slides from nonclinical species obtained during toxicology testing. Public partners such as hospitals and pathology laboratories are an invaluable source of clinical slides and associated data, from clinical trials, observational studies and archives.

Secondly, the infrastructure to host such collections can only be the result of the combined efforts by public and private sectors. Moreover, the interactions between academic, pharmaceutical industry and small and medium-sized enterprise (SME) partners will constitute a significant factor of success for the development of innovative software tools and efficient end-user applications. Lastly, the involvement of representatives of health and regulatory authorities will allow frameworks for policies or roadmaps pertaining to the validation and qualification of digital slides and their use for peer review, primary read and adjudication of nonclinical studies and clinical cases.

Scope

The overall scope of the Call topic is to collect, host and sustain virtual slides along with associated data and to support the collaborative development of artificial intelligence in pathology.

The funded action will also address the regulatory, legal and ethical challenges associated with the collection, sharing and mining of the virtual slides.

Objective 1: Sustainable infrastructure

To deliver the infrastructure hosting several petabytes of digital slides and making the data accessible for research. It represents the hardware layer of the funded action and could take the form of a data centre, either centralised or decentralised. The key factors of success for this objective are the storage capacity and the possibility to exchange rapidly large amounts of data.

The achievement of this objective is also critical for sustainability and the long-term impact of the funded action. The ambition is that after the end of the funded action, the repository will be maintained and developed, following a model similar to public repositories for genomics (e.g. National Center for Biotechnology Information (NCBI) /Gene Expression Omnibus (GEO) — https://www.ncbi.nlm.nih.gov/geo/) and that it becomes the central place for hosting raw digital slides associated with scientific and medical publications. The planned infrastructure is expected to allow pathologists to concomitantly review difficult cases and to consolidate large case series including histopathology and clinical information in order to establish diagnostic criteria. The sustainability beyond the end of the funded action will take the form of a business model that leaves open access free of charge for non-profit purposes. This will represent a major advantage compared to the current approach of smaller databases.

Objective 2: Data

To compile digital histopathology slides from nonclinical safety studies, as well as from clinical series needed to populate the initial version of the repository, and contribute to developing tools and artificial intelligence models. The key factor of success is the diversity of lesions, tissues, and species while providing sufficient sample sizes. In addition, the slides will be made publicly available for the development of artificial intelligence in pathology in line with the sustainability model described in objective 1.

Initiative Text

Objective 3: Tools

To deliver a mechanism of an honest broker (see 'Expected key deliverables' and 'Suggested architecture of the full proposal' sections) by developing a software ensuring the optimal and secure contribution of clinical and nonclinical material. Efforts will also be undertaken to propose a unified open digital slide format and tools to search, access, upload, register, download, view and homogeneously annotate information. In addition, AI models and tools, such as assistance to general diagnosis, screening for slides for lesions, and content-based image retrieval will be developed at a later stage of the funded action.

Objective 4: Regulatory framework

To advance the regulatory framework around the utilisation of digital pathology slides for nonclinical safety testing, evaluation of clinical trials and dissemination/discussion of difficult clinical cases. This will accelerate the adoption of roadmaps for the qualification of the usage of digital slides for peer-review or primary slide reading, as well for the development of artificial intelligence based tools for pre-screening and assisted diagnosis. This objective should be achieved by building on already existing and ongoing interactions and efforts between health and regulatory authorities, and professional societies.

Expected key deliverables

Based on these objectives, a number of key deliverables have been identified:

- mechanisms for adequate management of confidential information possibly associated with digital slides, through the establishment of a specific entity (further referred to as the honest broker);
- sustainable infrastructure to host a large series of digital slides (approximately three million during the lifetime of the project) ensuring confidentiality and privacy through the application of an honest broker concept. Meta-data and annotations will be provided in compliance with existing standards[1];
- nonclinical slide collection: approximately two million slides covering all tissues from several species and with the broadest spectrum of lesions should be collected. This material, obtained from toxicology studies, prospectively whenever possible, will represent a uniquely valuable asset for the fast development of models. Lesions elicited during toxicity testing are progressive and often in relatively pure form which is useful for developing models that recognise elementary lesions. Furthermore, such models developed initially on animal tissues can with little additional effort be expanded to clinical tissues and more complex lesions. It is required that the slides meet high standards of quality (e.g. orientation of samples, section thickness, staining) in order to optimally contribute to the development of AI models;
- clinical slide collection compliant with the quality and ethical standards: approximately one million digital slides should be provided from the archives and/or prospectively collected in the routine clinical practice over the project lifetime. They should be in a form of documented clinical series covering all the diseases areas such as (but not limited to):
    - oncology (e.g. breast, prostate and colon carcinoma, non-small cell and small cell carcinoma of the lung, hepatocellular carcinoma, or renal cell carcinoma, etc.);
    - dermatology (e.g. lupus, atopic dermatitis, melanocytic lesions, drug-induced skin reactions);
    - hepatology (e.g. autoimmune hepatitis, alcoholic and non-alcoholic steatohepatitis, drug-induced hepatitis, allograft rejection, tumours);
    - nephrology (e.g. glomerulonephritides, tubulointerstitial nephritides, drug-induced kidney injury, allograft rejection);
    - pneumology (e.g. idiopathic pulmonary fibrosis/usual interstitial pneumonia, nonspecific interstitial pneumonia).
- the established open-source data format for digital slides;
- developed open-source, cross-platform software tools to:
    - upload, search and access slides and associated metadata;
    - visualise and annotate the slides;
    - download slide for data mining and model development.
- AI models for:
    - identification of tissues and lesions;
    - generation of morphological and molecular signatures from slides.

[1] For example: International Harmonization of Nomenclature and Diagnostic Criteria (INHAND — https://www.toxpath.org/inhand.asp), Standardization for Exchange of Nonclinical Data (SEND — https://www.toxpath.org/send.asp) or International Classification of Diseases (ICD — https://www.who.int/classifications/icd/en/)

Indicative text

- engagement with regulatory authorities for adapting guidelines to the new field of digital pathology;
- a sustainability plan for the maintenance and future development of the repository towards a central place gathering virtual slides from clinical cases series and raw data associated with publications. The plan should explore and propose a business model making the use of digital slides for commercial developments subjected to fees, while open access for research purposes should remain free of charge. Besides funding the storage of a massive amount of slides, the plan should also include the activities related to the control of the high quality of slides and validation of new slides while enriching future collection.

Expected impact

Applicants should describe how the outputs of the project would contribute to the following impacts and include baseline, targets and metrics to measure impact:

- catalyse research in digital pathology by providing a unique combination of animal and human histopathology. By offering the first complete coverage of tissues and elementary lesions, this repository will offer an unprecedented opportunity to build holistic models and allow generic mining of histopathology, irrespective of a particular tissue or indication;
- enable the development of artificial intelligence tools for rare diseases and uncommon conditions, which currently are left out of the models because of the paucity of cases;
- help identify sub-types in common diseases, possibly unveiling new clinico-pathological entities amenable to specific therapeutic interventions. It could also contribute to assessing the translatability of animal models for disease modelling, safety and efficacy studies, and thereby rationalise the design of clinical trials and the use of animal models. Ultimately, it should accelerate and improve patient treatment and management, thereby enhancing patient health along with the more efficient use of healthcare resources;
- clear the way for the use of digital slides in nonclinical safety and clinical consultation, and facilitate the approval of AI-based tools for slide screening and assisted diagnosis;
- in the long term, the repository delivered by the consortium will be maintained through sustainability mechanisms defined by the consortium and will provide the community with an infrastructure to host additional digital slides (e.g. associated with the publication of case reports, cases series for disease stratification and clinical trials).

Applicants should indicate how their proposal will impact the competitiveness and industrial leadership of Europe by, for example engaging suitable SMEs.

Potential synergies with existing Consortia

Applicants should take into consideration, while preparing their short proposal, relevant national, European (both research projects as well as research infrastructure initiatives), and non-European initiatives. Synergies and complementarities should be considered in order to incorporate past achievements, available data and lessons learnt where possible, thus avoiding unnecessary overlap and duplication of efforts.

Therefore, the applicants should explore possibilities of synergies with a similar past and ongoing IMI1 and IMI2 as well as upcoming IMI2 projects.

Industry consortium

The industry consortium is composed of the following EFPIA companies:

- Novartis (Lead)
- Janssen (Co-lead)
- Bayer
- Boehringer Ingelheim
- Novo Nordisk
- Pfizer
- Roche
- Sanofi
- Servier
- UCB

The industry consortium will contribute the following expertise and assets:

Innovative Next

- the major part of the contribution will consist approximately in two million digital slides, mostly prospectively collected from high-quality nonclinical safety studies. These activities will be crucial to gather sufficient critical mass of high-quality slides needed for achieving the planned objectives;
- digital slides from clinical trials will be brought in. However, the vast majority of the clinical collection will be provided by the applicant consortium (see work package 3 'expected applicant consortium contribution');
- experience and guidance for the harmonisation of metadata associated with digital slides;
- experience and guidance for the interaction with health authorities with respect to the qualification of digital and computational pathology in drug development.

Indicative duration of the action

The indicative duration of the action is 72 months.

Indicative budget

The indicative in-kind and financial contribution from EFPIA partners is EUR 37 771 260.

Due to the global nature of the participating industry partners, it is anticipated that some elements of the contributions will be non-EU/H2020 Associated Countries in-kind contributions.

The financial contribution from IMI2 JU is a maximum of EUR 32 320 000.

Applicant consortium

The applicant consortium will be selected on the basis of the submitted short proposals and it is expected to address all the objectives and make key contributions to the defined deliverables in synergy with the industry consortium which will join the selected applicant consortium in preparation of the full proposal for stage 2.

This may require mobilising, as appropriate the following expertise and capabilities:

- proven expertise in the management of digital slides in various formats including mastering of tools/mechanisms to collect/extract digital slides from various places (e.g. sponsors, contract research organisations (CROs)), transferring them securely to a central repository, and ensuring derived data can be returned to the contributor on demand;
- expertise in developing large databases for digital slides and related metadata, and tools to interact with them. Metadata correspond to various modalities associated with digital slides accessible for example via clinical registries, electronic health records, e.g. tabulated summaries of elementary lesions for non-clinical toxicology studies, summaries of information on the diagnosis, clinical presentation, genetic abnormalities and/or biomarker values for clinical samples;
- expertise in developing end-user applications for the visualisation, annotation, and analysis of digital slides;
- expertise in managing large clinical databases and large amounts of data;
- proven mastering of methodologies in creating tools for editing labels, anonymising/coding digital slides, encrypting individual files, and other methodologies required to set up the mechanism of the honest broker;
- the expertise of developing and training large-scale deep learning models for histopathology, such as convolutional neural networks, and evaluating the performance thereof;
- expertise in generating, annotating and sharing digital slides;
- solid scientific, medical, and clinical (including pathologist) expertise and knowledge in the research areas targeted by the topic text;
- legal, ethical and regulatory expertise related to patient privacy, informed consent, data anonymisation, and electronic submission of trial/safety data;
- professional project data management and communication capabilities with previous experience in large European public-private partnership settings.

In their proposal, applicants should demonstrate access to the following resources:

- proven access to large and well clinically documented collections of digital slides from clinical and diagnostic cases (e.g. from well-established pathology department(s)) relevant to disease areas enumerated under 'Key deliverables', organised in series with appropriate informed consent and

Informative text

preferred molecular biomarker annotation (e.g. next generation sequencing (NGS) oncogene panels or whole exome sequencing);
- adequate infrastructure and computing power to train deep-learning models, host and make accessible large amounts of data (approximately 3 peta-bytes for three million digital slides);
- infrastructure to scan a large number of slides (approximately one million).

Suitable SMEs can, for instance, be considered for the following activities: infrastructure management, honest broker mechanism, end-user interfaces and slide scanning.

The suggested architecture of the full proposal

The applicant consortium should submit a short proposal which includes their suggestions for creating a full proposal architecture, taking into consideration the industry participation including their contributions and expertise provided below.

In the spirit of the partnership, and to reflect how IMI2 JU call topics are built on identified scientific priorities agreed together with EFPIA beneficiaries/large industrial beneficiaries, these beneficiaries intend to significantly contribute to the programme and project leadership as well as project financial management. The final architecture of the full proposal will be defined by the participants in compliance with the IMI2 JU rules and with a view to the achievement of the project objectives. The allocation of a leading role within the consortium will be discussed in the course of the drafting of the full proposal to be submitted at stage 2. To facilitate the formation of the final consortium, until the roles are formally appointed through the consortium agreement, the proposed project leader from among EFPIA beneficiaries/large industrial beneficiaries shall facilitate an efficient negotiation of project content and required agreements. All beneficiaries are encouraged to discuss the project architecture and governance and the weighting of responsibilities and priorities therein.

The proposal should be articulated around the following phases, which may overlap as needed to allow the optimal utilisation of resources and production of deliverables:

Phase 1: Establish an honest broker and infrastructure.

Phase 2: Data collection, tools for access and visualisation.

Phase 3: Artificial intelligence models and tools for morphological data mining and assisted diagnosis.

The architecture outlined below for the full proposal is a suggestion. The architecture of the full proposal should be designed to fulfil the objectives and key deliverables within the scope of this topic.

Work package 1 – Project management, coordination, and sustainability

This work package will address the strategy and implementation of project management. This will encourage regular meetings and interaction between sub-groups and teams to coordinate and follow up on the work effort. The applicant consortium with input from industry partners will develop the sustainability plan. Its objective should be to provide an infrastructure to host additional digital slides contributed by authors of case reports, clinical series or clinical trials, with the same level of annotation, anonymisation and accessibility for model development, as during the research phase. The plan should comprise financial, legal, ethical and structural aspects as well as scalability of the storage/access capacity.

Industry contribution:

Assurance of the coherence of consortium activity, and involvement in project management including planning, budgeting, follow-up and tracking of the work packages' progress, and consolidation of the reports. Project risk management and comprehensive communication and dissemination of the project's progress and its milestones will also be provided.

Expected applicant consortium contribution:

Providing detailed follow-up and tracking, via regular work package reports, early reports of any unexpected organisational or structural issues or delays with respect to the project deployment and intermediate objectives.

Initiative text

Work package 2 – Infrastructure and database hosting

This work package consists of the development of the infrastructure that will host approximately three million digital slides shared during the course of this project, and ensure that they are easily accessible to other project participants through available internet servers. The applicant consortium will ensure that the proposed infrastructure is amenable to expansion and is coordinated with the sustainability plans. The choice of the infrastructure will be coordinated with the industry partners and other consortium partners to ensure compatibility with the tools.

Industry contribution:

Advice for the harmonisation of metadata associated with the digital slides provided.

Expected applicant consortium contribution:

Building an infrastructure (data centre) to host three million digital slides and implement a database to register the corresponding files and associated metadata.

Work package 3 – Data collection & management

To support the other work packages, a data management system/database, able to register the digital slides contributed to by the industry partners and the applicant consortium, is needed. It will ensure the encoding of the data and compliance with patient privacy legislation and the confidentiality agreements established with the industry partners through an honest broker mechanism. The data management should also ensure that contributed digital slides, stripped from all proprietary information, are coded while retaining links with associated metadata (e.g. species, staining, tissue), and possibly complementary data such as clinical pathology, biomarkers, omics profiles, when shared by the contributor. Metadata will use controlled terms from the International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) or International Classification of Diseases (ICD) classifications. This work package also comprises the handling, shipping and scanning of cases contributed as glass slides.

Slide scanners currently deliver the file in a proprietary format, which has limited compatibility outside the product family. In addition to data management, this work package will deliver a common, unique file format for virtual slides that are compatible with open-source visualisation software, where images associated with the virtual slide such as the label or the overview can be edited in order to remove confidential information.

Industry contribution:

Approximately two million glass or digital slides from nonclinical toxicology studies, animal models of diseases, or clinical trials, along with metadata, compliant with INHAND/ICD nomenclature, whenever possible, and structured under the standardisation for exchange of nonclinical data (SEND) format.

Expected applicant consortium contribution:

- honest broker mechanism: to allow all participants to share data comfortably in a secure environment, the applicant consortium should include an organisation with a proven track record of acting as an independent honest data broker from a legal and historical perspective. The mechanism and expected contribution should consist of:
  - setting up the database, encoding mechanisms and registering digital slides accordingly;
  - ensuring that digital slides contributed by members of the consortium are stripped from any information that could link them back to a specific study or patient when made available for the project (including elements of the digital slides themselves such as pictures of the original label);
  - ensuring information security and managing access rights between members of the consortium and the public, at the level of the individual digital slides through encryption;
  - keeping the possibility for a contributor to link scientific results (e.g. model predictions) to the contributed slide, if requested at the time of the submission of the digital slide;
  - if glass slides are submitted, organising their physical transfer to scanning facility, registration in the repository and return to the contributor.

Indicative text

- digital or glass slides from clinical series and archives: the clinical partners of the applicant consortium will provide approximately one million digital or glass slides from clinical case series obtained from the archives and/or prospectively collected from routine clinical practice in pathology laboratories, with accompanying diagnostic and clinical data using a controlled vocabulary (e.g. ICD);
- scanning of glass slides.

Work package 4 – Tools for accessing, annotating and mining digital slides

This work package intends to develop the following tools:

- tools for accessing slides: software tools to interact with the database will be developed to enable access to the virtual slides and the related metadata through search functionalities;
- tools for visualisation and annotation: the annotation of virtual slides refers to the delineation of regions of interest representing particular tissues, features, structures or lesions. Currently, available tools offer some of the required functionalities, which are usually insufficient to perform complex annotation tasks required for the training of deep-learning based models. Cross-platform, open-source tools will be developed to visualise and navigate fluently virtual slides of various file formats hosted in the database, including possible original formats developed in this project. The software tool will offer annotation functionalities for the optimal annotation of slides by pathologists and histologists.

Industry contribution:

- defining the functionalities required;
- guiding the development of tools to ensure implementation according to required functionalities;
- testing tools and providing feedback.

Expected applicant consortium contribution:

- providing tools to interact with said databases and managing metadata along with the digital slides;
- setting up end-user applications for the visualisation, annotation, and analysis of digital slides;
- providing large-scale deep learning models for histopathology, such as convolutional neural networks.

Work package 5 – Regulatory framework for digital slides and AI-based methods

The consortium is expected to have a strategy for the translation of the relevant project outputs such as policies or frameworks for the qualification of the use of digital pathology slides for peer-review and primary reading in nonclinical safety assessment and evaluation of clinical efficacy. It will explore the optimal utilisation of the digital slides from patients to develop AI in pathology in compliance with the General Data Protection Regulation (GDPR). It will also envisage the roadmap for the qualification of AI-based tools for the pre-screening of normal tissues in nonclinical safety and possibly selected domains of clinical practice. A plan for interactions with regulatory agencies/health technology assessment bodies with relevant milestones and allocated resources should be proposed to ensure that at least qualification advice or opinions are provided on the proposed methods during the course of the funded action.

Use of digital slides: the project will provide a platform to exchange and publish virtual slides from nonclinical and clinical studies. Although professional associations and some regulatory bodies have already developed guidance or opinions regarding the use of digital pathology techniques for regulated laboratory work, their applicability is still limited. This project will ideally accelerate the dialogue and create an interface between health authorities, regulatory bodies, clinicians and the pharmaceutical industry regarding the use of digital slides for the primary assessment of nonclinical safety studies, clinical trials and diagnosis.

AI-based methods: the ambition of the project generated from this topic is to catalyse the development of artificial intelligence in pathology by facilitating access to digital slides, a critical resource for training deep-learning based models. These models could serve as prediction engines for assisted diagnostics tools. This project should provide a platform for interaction between the scientific experts and health authorities aiming towards defining a framework for the qualification of these complex tools for clinical and regulatory use, e.g. the project's central repository could be used as a clinical reference or external quality assessment tool for pathologists.

Indicative text

Industry contribution:

Guidance for the interaction with health authorities with respect to the qualification of digital and computational pathology in drug development.

Expected applicant consortium contribution:

- engaging with health authorities representatives to get input to be discussed in the different governance structures of the funded action;
- organising and leading discussions for the adoption of frameworks or roadmaps for the qualification of the usage of digital slides and AI tools as described in the topic text, the use of clinical slides from archives and for the sharing of rare cases or published cases series. Therefore, the overall contribution should consist of:
    - contribute to the evolution of the use of digital slides as a surrogate of glass slides in nonclinical safety assessment (peer-review, primary read);
    - establishing a framework for or a roadmap towards the validation/qualification of artificial intelligence tools for nonclinical safety applications such as screening, lesion detection and grading, and for routine clinical use such as support for lesion detection, qualification/quantification of events, clinical decision-making support tools;
    - contribute to the evolution of the regulatory framework around the use of clinical slides from archives and AI tools in clinical trials;
    - defining the regulatory context for the sharing of rare cases or published cases series.

Indicative text