

BBMRI-ERIC: How is Medical Data Used for Research?



BBMRI-ERIC®
biobanking for a healthier world

Assoc. Prof. Petr Holub, Ph.D.
Chief Information Officer (CIO)

BBMRI IS A DISTRIBUTED ERIC

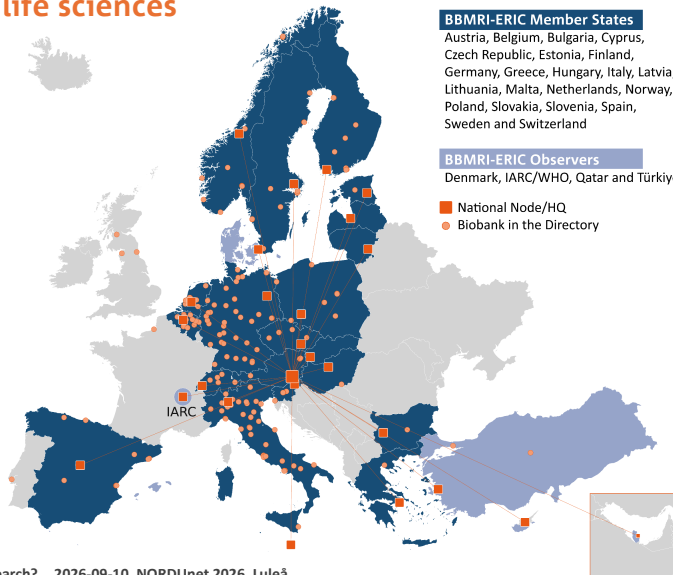
Operating in health and life sciences



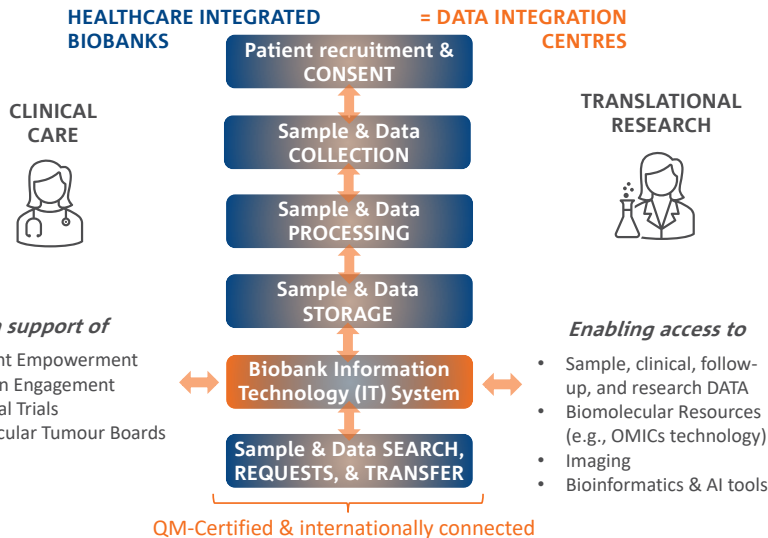
THE BBMRI COMMUNITY

~ 500 biobanks at clinics and universities across 32 countries
25 National Nodes + IARC/WHO
3 Biotechnology Expert Centres
32 ELSI experts
130 Information Technology experts
313 Quality Management experts
1 Headquarter

SERVING SOCIETY,
ACADEMIA AND
INDUSTRY ALSO AS DATA
INFRASTRUCTURE



THE LOCAL BIOBANK ECOSYSTEM



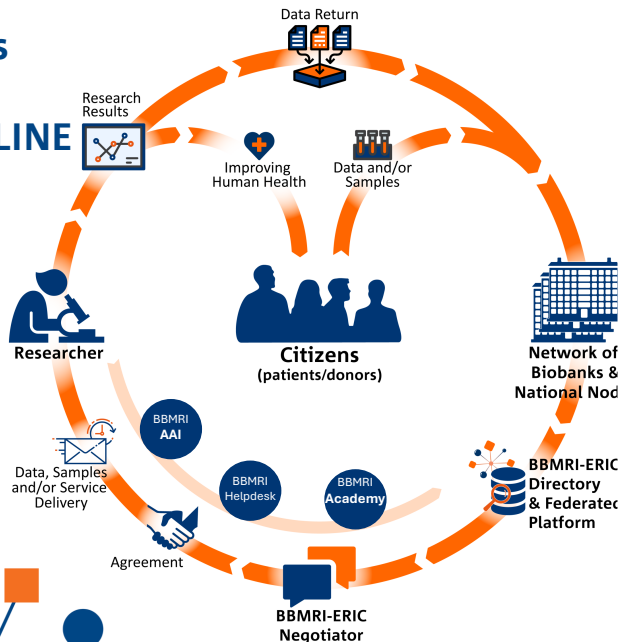
BBMRI-ERIC's FEDERATED ACCESS PIPELINE

BBMRI-ERIC Directory is largest biobank catalogue worldwide:

- ✓ 473 biobanks with
- ✓ 9.9 Mio. citizens across
- ✓ 33 countries.

BBMRI-ERIC Federated Platform features:

- ✓ 450,000 citizens with
- ✓ 155,000 genomes across
- ✓ 10 EU countries today.



Selected KPIs 2024 (2023)

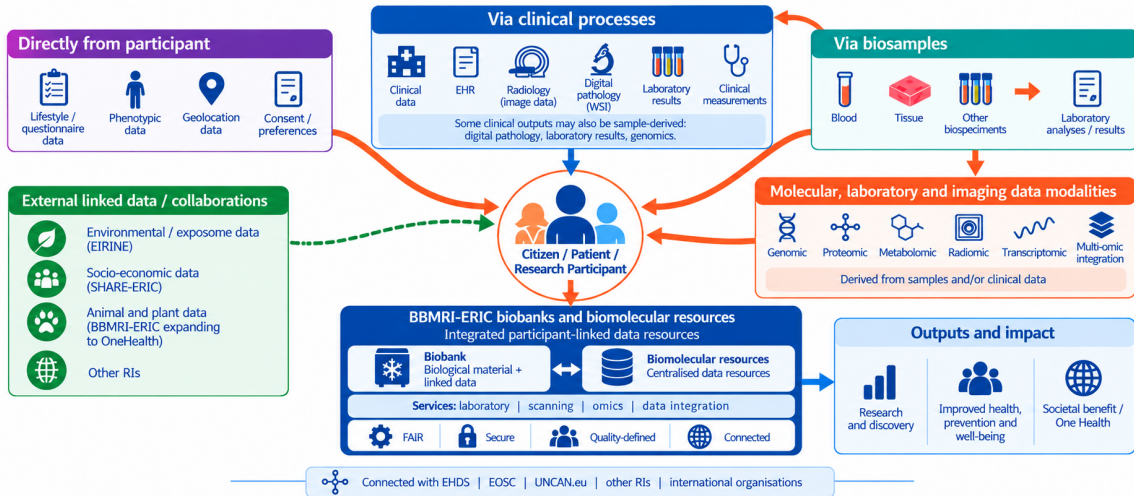
10,444 (8,924)
users of the
Directory

9,134 (6,392)
successful
access requests

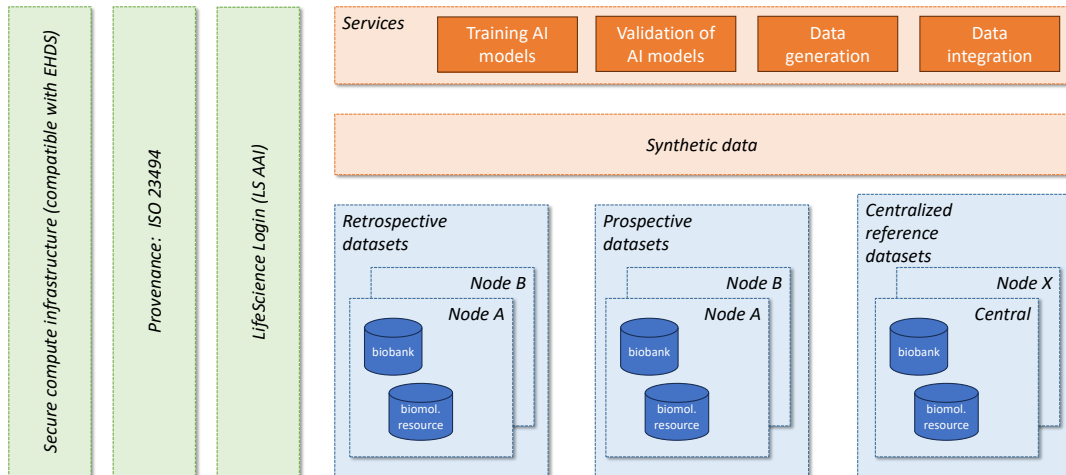
3,204 (2,566)
publications
involving BBMRI

42,992 (24,807)
participants in
outreach events

Data Modalities in BBMRI-ERIC



Data-Centric Service Portfolio of BBMRI-ERIC



Access To What?

▶ **Samples and/or Pseudonymized data**

- access requested directly via a biobank or via **BBMRI-ERIC Negotiator** (connected to all discovery services)
- authenticated users under contract (MTA/DTA)
- data transfer or data provisioning via SPEs

▶ **Anonymized data**

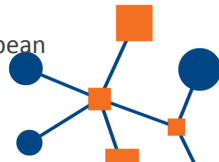
- authenticated users after agreeing terms and conditions

▶ **Synthetic data (under development)**

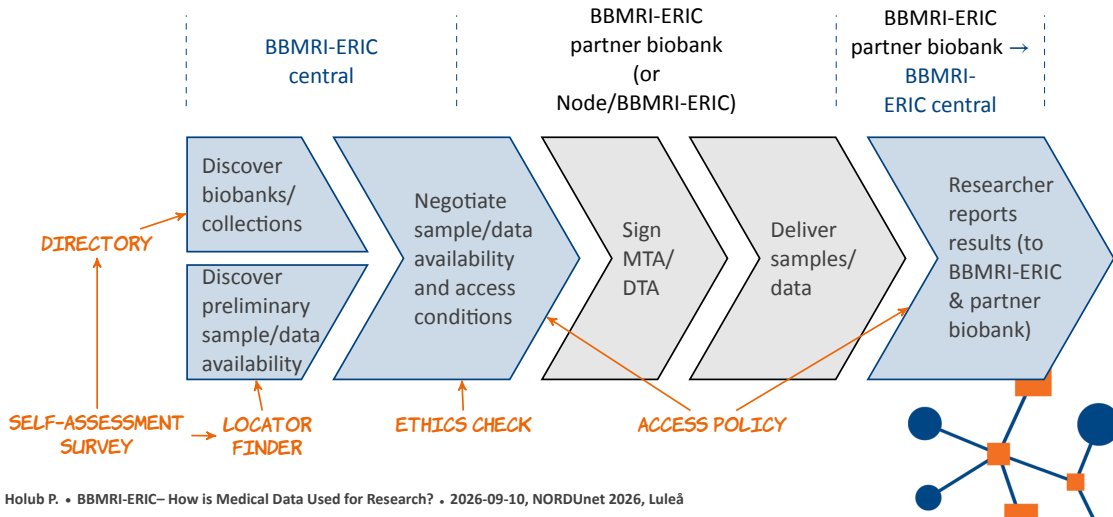
- temporary proxy data before actual access can be granted

▶ **Services**

- access requested directly, via BBMRI-ERIC Negotiator, or via European Trans-National Access (TNA) projects such as canSERV



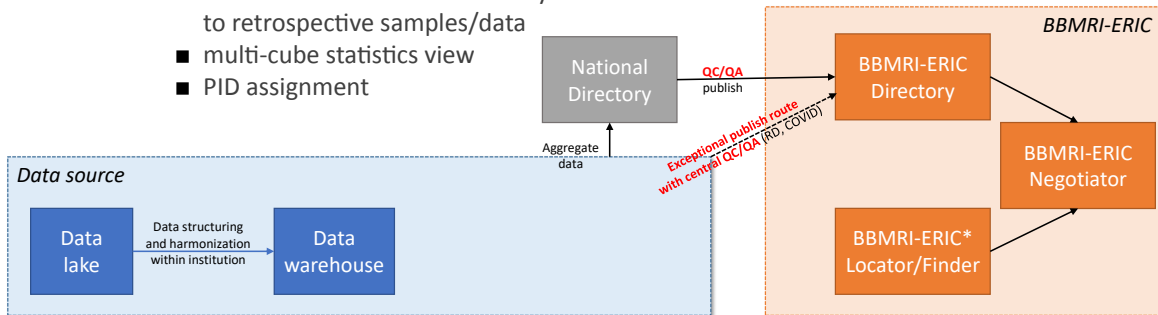
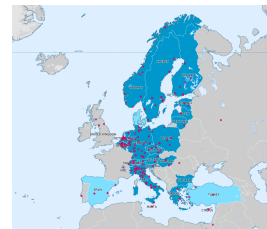
BBMRI-ERIC Access Pipeline



BBMRI-ERIC Federated Data Access

► Meta-data catalog only findability

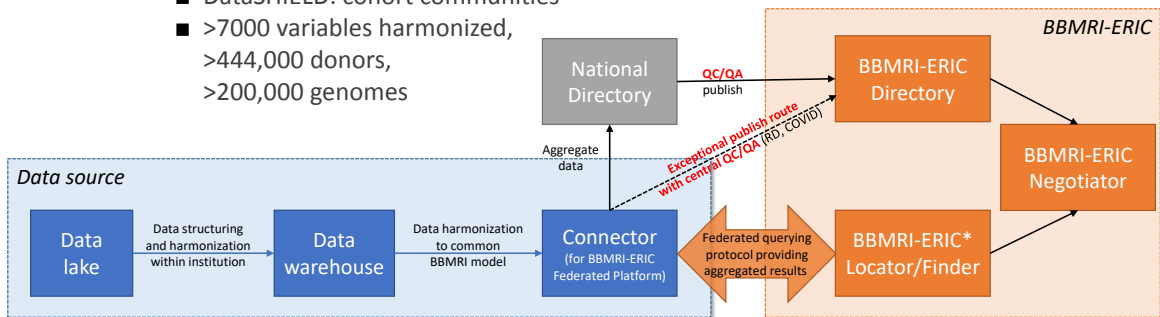
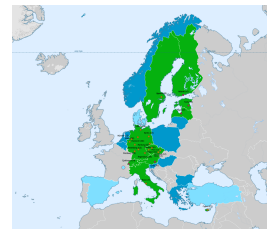
- Directory: providing aggregate descriptors on all BBMRI-ERIC biobanks and biomolecular resources
- can advertise other services beyond access to retrospective samples/data
- multi-cube statistics view
- PID assignment



BBMRI-ERIC Federated Data Access

► Federated querying findability

- Locator & Finder: privacy-preserving availability querying and analyses (e.g., GWAS, PheWAS)
- DataSHIELD: cohort communities
- >7000 variables harmonized,
>444,000 donors,
>200,000 genomes

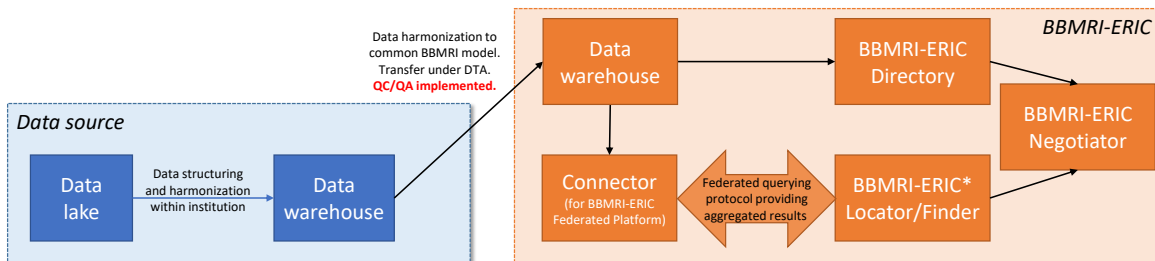
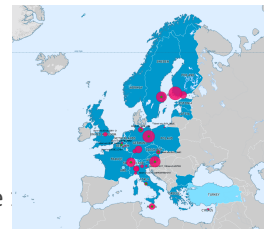


* Locator/Finder can have also national instance relaying to central BBMRI-ERIC

BBMRI-ERIC Centralized Data Access

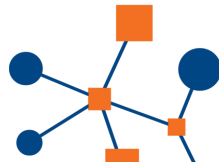
► Centralized data collections

- curated reference databases
- focus on efficient access and low barriers of access – biobanks have veto period
- e.g., CRC-Cohort, ongoing building of digital pathology imaging



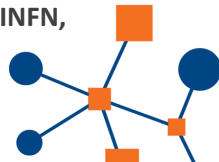
Access via Provisioning of SPEs/TREs

- ▶ **Data transfers pose increased security risks**
- ▶ **Depositing data into "Secure/Trusted Processing Environments"** where data can be accessed and processed but only aggregated/anonymized results retrieved
 - **optimization of functionalities, e.g., w.r.t. digital pathology and (X)AI**
- ▶ Supported via **RI-SCALE: Data Exploration Platforms (DEPs)**
- ▶ Alignment towards **European Health Data Spaces**



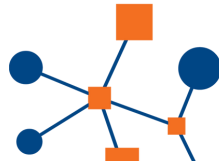
BBMRI-ERIC & eInfrastructures

- ▶ **BBMRI-ERIC is a Research Infrastructure – not an eInfrastructure**
 - overall, BBMRI-ERIC is not systematically building compute/storage capacities, but rather builds on eInfrastructures
 - some notable exceptions: NTNU/HUNTCloud, UMCG
- ▶ **Areas of collaboration:**
 - storage/compute for biobanks and biomolecular resources
 - also to help connect to the BBMRI-ERIC infrastructure
 - storage/compute for central BBMRI-ERIC and its Nodes for data pooling
 - providing trustworthy TREs/SPEs to the users of biobanks and biomolecular resources
 - ... which data sources need to be able to trust
- ▶ **More systematic collaboration within EOSC: e.g., EOSC-IT/ICSC/INFN, EOSC-CZ/MUNI/SensitiveCloud, EOSC-NL/SURF**
- ▶ **Other collaborations: MUSICA (AT)**



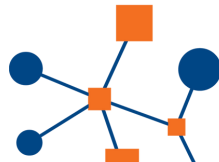
BBMRI-ERIC & Research use

- ▶ **Research with direct clinical impact**
 - improving **diagnosis** (biomarkers) and **treatment** of diseases
 - improving **population health** – preventive programs, early detection of problems
- ▶ **Biological research – precursor to clinical impact**
 - elucidating biological mechanisms of diseases
- ▶ **AI research – both basic research and applied clinical research**
- ▶ **Providing portfolio mix: data – samples – services – expertise**
- ▶ **2,566 (2023) → 3,204 (2024) → 4,568 (2025) publications**

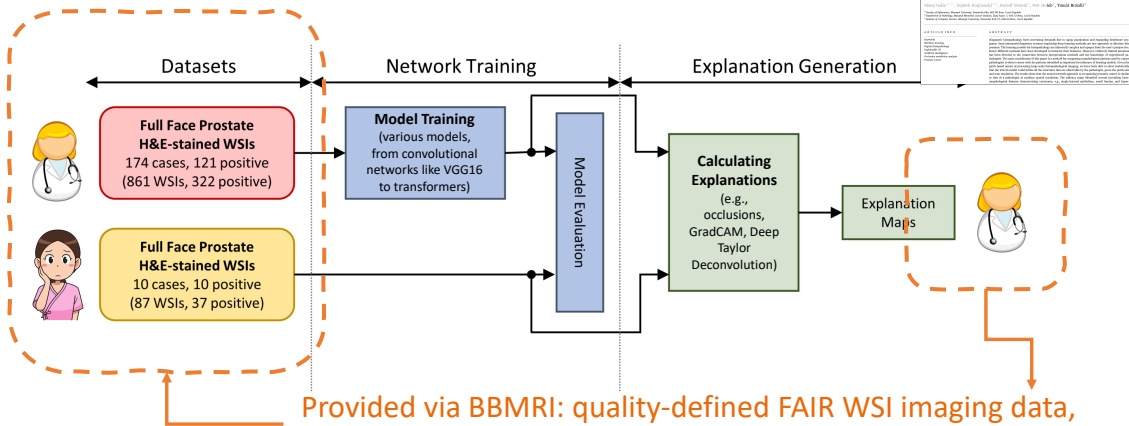


BBMRI-ERIC & AI

- ▶ **BBMRI-ERIC as a data provider for clinically-relevant AI**
 - sample-derived data (imaging, omics, ...) and associated data (clinical, phenotype, exposure, ...) for **model training**
 - environment to **validate AI models** on unseen data – **single-centric** or **multi-centric**
- ▶ **BBMRI-ERIC as an expertise provider for clinically-relevant AI**
 - identification of clinical experts
- ▶ **BBMRI-ERIC as a host of AI groups – to improve services of the RI, e.g.,**
 - **quality checking**
 - **providing descriptors** for data sets **to facilitate discovery**
 - **generate synthetic data** to speed up or proxy access

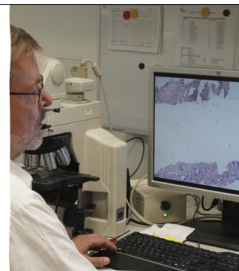
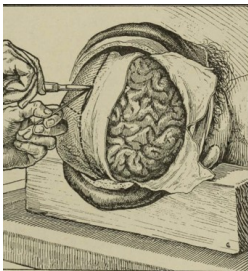


How is BBMRI-ERIC used for XAI?



MCVAL EOSC Federation Use Case

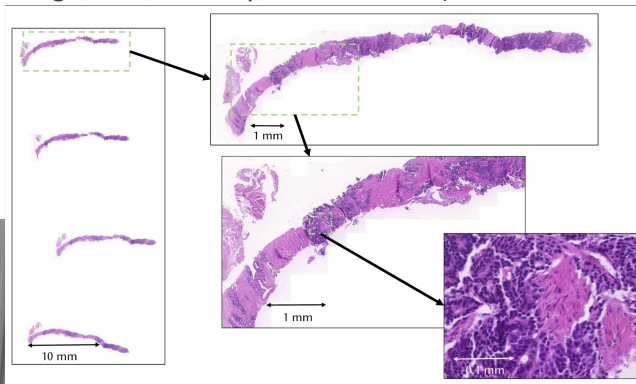
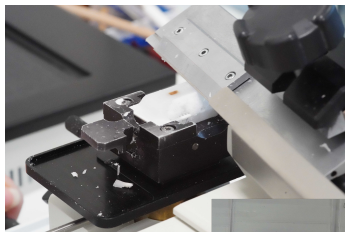
- ▶ **Evolution of pathology:** from examining dead bodies to **critical cancer diagnostics**
- ▶ Projected digital pathology growth: \$1.2B (2024) to \$2.6B (2032) (Delvens [↗](#))



MCVAL EOSC Federation Use Case

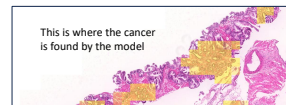
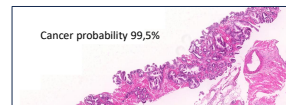
Example of prostate slide

- Statistically: about 1 in 8 men will be diagnosed with prostate cancer during their lifetime (risk varies with age, race/ethnicity, other factors...).



MCVAL EOSC Federation Use Case

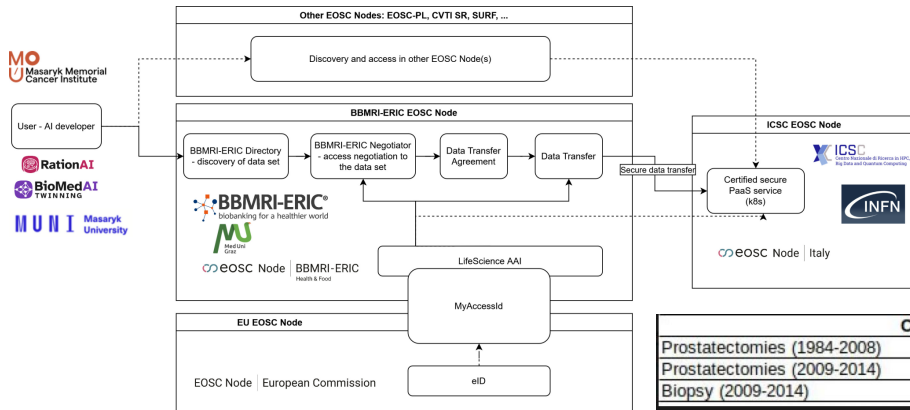
- **Validation of prostate cancer detection model developed and used by Masaryk Memorial Cancer Center (BBMRI.cz host) and Masaryk University, on multi-centric European data**
 - in routine clinical use at MMCI after internal validation
 - significant exploration of explainability of the AI model
 - BBMRI can bring in data sources and scientific communities to EOSC
 - initially piloted with BBMRI.at data
 - successfully **demonstrated during EOSC Symposium 2025**



EOSC Multi-Centric AI Validation Demonstrator

Roles – 36 experts involved!

- AI researcher
- Secure compute node
- Sensitive data provider
- AI validation service provider



Participants

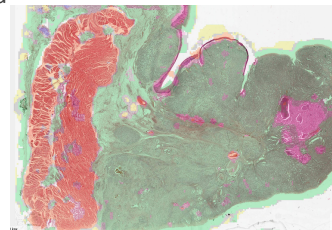
- MMCI/MU
- ICSC/INFN
- MUG
- BBMRI-ERIC

Additional EOSC Federation Sensitive Data WG Participants

- EOSC-SK
- EOSC-PL
- EOSC-NL (SURF)
- EOSC-FI

	Cases	Slides
Prostatectomies (1984-2008)	5525	104335
Prostatectomies (2009-2014)	794	28.786
Biopsy (2009-2014)	331	11.193

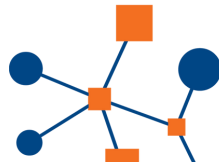
- improving **findability in large archives of WSIs**
 - only Medical University Graz and Masaryk Memorial Cancer Institute have cca 3 PB (>500k) WSIs
 - archives with 50M+ histopathological slides could be scanned
- **rough classification of tissue types and organs**
on otherwise undocumented slides
- **generating metadata descriptors** to enable findability
- steer synthetic image generation
- <https://doi.org/10.1016/j.artmed.2026.103368>



Data Anonymization and Synthesis

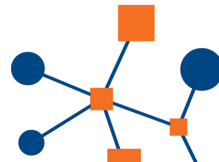
Having easy access would be largely beneficial – everybody is now using TCGA...

- ▶ **How can we streamline access to data?**
- ▶ **What is actually the risk of considering WSIs anonymous?**
- ▶ **Can we provide access to “conditioned” synthetic data as a temporary proxy to real data?**
- ▶ **How can we make user interfaces for access more efficient?**



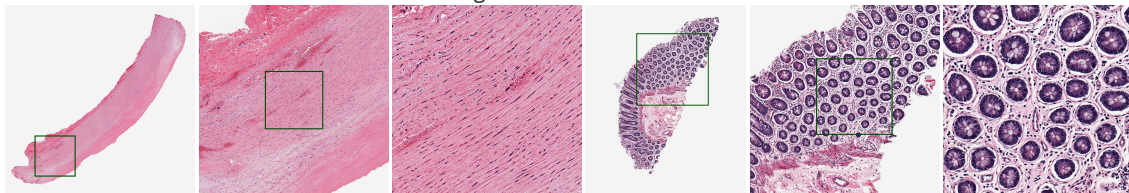
-

- <https://www.nature.com/articles/s41467-023-37991-y>



Data Anonymization and Synthesis

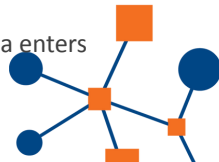
- ▶ **Ability to fundamentally accelerate/pipeline access if synthetic data is available**
 - generating temporary proxy data on demand to match real data
 - preserving privacy while synthesizing data
- ▶ **Recent work on gigapixel WSI data synthesis by Dr. Müller Group (MUG) – lead by Robert Harb**
 - multi-scale diffusion-based generative model



<https://doi.org/10.48550/arXiv.2311.08199>

How EHDS Changes the Landscape (1)

- ▶ **European Health Data Spaces (EHDS) is an alternative access pathway – using EU Regulation**
 - does **not supersede existing/working mechanisms**
 - **helps to unblock data** – blocked otherwise because of missing legal basis (e.g., limited consent), competing interests, lack of collaboration, vertically partitioned data across different data holders (**Nordic!**), etc.
 - very **powerful instrument** – hence coming **with limitations**: only possible to process data in SPEs for limited period of time and retrieve anonymized results
 - **not suitable for some purposes** – e.g., building permanent international patient/data cohorts
 - does not solve legal basis for collection and processing before data enters EHDS



How EHDS Changes the Landscape (2)

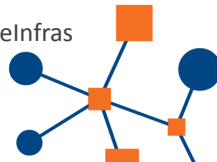
► **BBMRI-ERIC**

- central: authorized participant, possibly also a (trusted) data holder (TDH) for centralized collections
- biobanks and biomolecular resources: mostly (trusted) data holders
- National Nodes: various roles depending on the node setup

► **BBMRI-ERIC was involved in EHDS since 2020 and now supporting Node (per request) in their national EHDS adoptions**

► **Working with eInfrastructures on implementing SPEs**

- national/international SPEs
- each TDH is expected to provide and SPE – possibly via service of eInfras



THANK YOU!



BBMRI-ERIC[®]
biobanking for a healthier world

petr.holub@bbmri-eric.eu

Research Infrastructures and Data Sharing Infrastructures

