Roadmap for a European cancer data management and precision medicine infrastructure

Macha Nikolski, Eivind Hovig, Fatima Al-Shahrour, Niklas Blomberg, Serena Scollen, Alfonso Valencia, Gary Saunders

Roadmap for a European Cancer Data Management and Precision Medicine Infrastructure

Macha Nikolski[1,2,*], Eivind Hovig[3,4,*], Fatima Al-Shahrour[5,*], ELIXIR Cancer Data Focus Group, Niklas Blomberg[6], Serena Scollen[6], Alfonso Valencia[7,8], Gary Saunders[9]

1. Univ. Bordeaux, CNRS – IBGC, UMR 5095, Bordeaux, France
2. Univ. Bordeaux, Bordeaux Bioinformatics Center CBiB, Bordeaux, France
3. Centre for Bioinformatics. Dept. of Informatics, Univ. of Oslo, Norway
4. Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Norway
5. Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
6. ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK
7. Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain
8. ICREA, Barcelona, 08010, Spain
9. ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK (current address: EATRIS-ERIC, Amsterdam, Netherlands)

* These authors contributed equally

Standfirst

Gold standard cancer data management is pivotal to enable precision medicine for European citizens. Achieving this goal relies on key elements: adopting standardized data formats, ensuring robust data privacy, educating professionals about the infrastructure's benefits, and leveraging cutting-edge technologies to transform cancer care.

Introduction

Precision cancer medicine thrives on understanding the biology of each patient's cancer and on the integration of data from different sources and experiments, a cornerstone for driving effective and cost-efficient treatment programs (Box 1). Reusing of patient data for research, known as secondary use of data, presents a considerable opportunity for personalised cancer medicine. This approach leverages cost efficiency, larger sample sizes, and the ability to form hypotheses, but necessitates vigilance regarding biases, data quality, and ethical considerations. While such data reuse in clinics is less frequent, it translates in data-driven recommendations for individual patients and facilitates clinical decision-making. Evaluating this secondary data within a clinically relevant timeframe for each patient is paramount, alongside aligning it with global research knowledge and best practices. However, this comes with practical challenges, notably in integrating and analysing these data effectively (Figure 1).

The Europe's Beating Cancer Plan [1], an initiative by the European Union that aims at reducing cancer morbidity on a pan-European level, introduces 13 recommendations. Notably, the 10th recommendation focuses on creating a network of infrastructures supporting cancer research and care [2]. To realise this, it's essential to to bring together the service offerings of European Research Infrastructures such as e.g. the European Research Infrastructure for biological data (ELIXIR) and the European Infrastructure for Translational Medicine (EATRIS), pan-European cancer-focused organisations like Cancer Core Europe [3], and internationally recognized standards setting agencies such as the Global Alliance for Genomics and Health (GA4GH). This convergence of efforts aims to better support research infrastructures in addressing the common challenges faced by cancer data researchers.

Facilitating interoperability of cancer data

Broad interoperability of data, metadata, research software and computational infrastructure is critical for capitalising on the potential of extensive datasets. European Infrastructures play a crucial role in facilitating data interoperability and advancing cancer research, with the ultimate goal of to improved patient care and treatment outcomes.

ELIXIR coordinates experts from European national nodes to set standards and tools for integrating diverse datasets across Europe. The ELIXIR Cancer Data Focus Group, comprised of cancer specialists within the ELIXIR community, specifically addresses European cancer data management and precision medicine strategies, leveraging ELIXIR's broader genomics data structure. The group works alongside the 1+Million Genomes (1+MG) initiative and key EU projects that support it, such as the Beyond 1 Million Genomes project (B1MG) and the European Genomic Data Infrastructure project (GDI), where cancer is a use case. Guidelines and solutions produced for secure access to genomic and clinical data across Europe are applicable for cancer data (see Figure 1 for an overview of datatypes), as well as the genomics data infrastructure.

This coordinated effort involves alignment with major European Infrastructure projects, such as EATRIS, the European biobanking research infrastructure (BBMRI), and the European Clinical Research Infrastructure Network (ECRIN) (Figure 2). All are involved in cancer data interoperability and are members of the research infrastructures' Working Group of the European Open Science Cloud (EOSC), a critical platform supporting cancer research by offering an open and federated environment for research data hosting and processing. ELIXIR, EATRIS, BBMRI, and ECRIN actively participate in the EOSC4Cancer project, aiming to enhance accessibility and interoperability of cancer data for researchers. Their involvement ensures the cataloguing of services in platforms like the EOSC Portal, ensuring their relevance, usability, and interoperability.

All these alignment efforts guarantee that these infrastructures are interoperable from inception, enabling accessible analysis of cancer data through platforms like the future European Health Data Space (EHDS).

Infrastructure challenges to the analysis of cancer data at the pan-European level

Analysing cancer data across Europe presents various infrastructure challenges, encompassing complexities in data collection, standardisation, storage, sharing, and analysis. Here are key challenges in analysing cancer data at the pan-European level:

Data Governance

Patient derived data management and access are critical in primary research collaborations and secondary analysis by the wider research community. However, challenges related to data governance, ethics, legal considerations, and patient consent hinder responsible and legitimate data sharing. A recent article highlighted the tension between open data access for secondary use and the need for appropriate data use [4]. It recommended five conditions for better governance of human genomic data: enabling access, complying with laws and agreements, ensuring appropriate use and minimising potential harms, promoting equitable data access and usage, and utilising genomic data for public benefit.

Addressing the data sharing and access challenge across Europe requires connecting national and transborder infrastructures, developing interoperable technologies, and converging towards infrastructures that adhere to common standards for secure and ethical sharing practices on a pan-European scale. Such data infrastructures will play a pivotal role in supporting and shaping European policy that govern health data sharing, access, and

utilisation, respecting cancer patients' rights while enabling controlled data access across jurisdictions. An emerging model, the Federated European Genome-Phenome Archive (EGA), maintains data locally while sharing metadata globally [5] using a federated data governance approach. The Federated EGA adopts GA4GH standards for service interoperability across different implementations and within the wider landscape of cancer data management infrastructures.

Data privacy and consent

The processing of personal data is generally prohibited unless explicitly permitted by law or participant consent. The GA4GH has highlighted several obstacles to secondary clinical data use, necessitating multidisciplinary collaboration [6]. Among them is consent that is insufficient for data sharing and reuse; indeed, consent must be given to conduct specific research, which forbids data reuse in other contexts. Balancing data availability and protection is crucial, notably in Europe, regulated by the General Data Protection Regulation (GDPR). The GDPR defines health data as information about a person's physical or mental health, mandating responsible data sharing through minimisation, pseudonymisation, and anonymisation.

European legislation shows significant variation in local anonymisation standards. Some jurisdictions demand nearly impossible re-identification, while others accept anonymisation if re-identification is unlikely [7]. For instance, France and Finland impose stricter de-identification norms than recommended by the European Data Protection Board (EDPB) under GDPR. Finland only permits public-access data aggregation; other health data, like pseudonymised data, requires access in secure monitored environments. France mandates a certified “Hébergeur de données de santé” for storing, analysing, and sharing health data.

Omics data, such as genome sequences, are notoriously difficult to anonymise, and multiple modalities of cancer data increase the possibility of tracing back to the individual. The PHG Foundation recent report identified GDPR-related challenges in genomic research, including data control, sharing limitations, accessibility, minimisation, and storage restrictions [8]. Genetic and health data fall under the category of sensitive or "special categories of personal data" under Article 9(4), allowing member states to introduce additional regulations. This leads to divergent national rules affecting data processing, sharing, and accessibility. The variance in regulations emphasises the necessity for federated solutions across Europe.

Cancer Data Multi-Modalities

Characterisation of cancer biosamples has generated petabytes of omics, imaging and associated clinical data [9]. However, accessing public research data repositories remains difficult due to several challenges, such as data location, data type compatibility, and data quality. Moreover, access to the accompanying clinical data requires approval and compliance, often leading to delays and additional costs. An additional infrastructure challenge is to provide efficient solutions for domain-specific challenges that may be specific to tumor types and subtypes, as well as for molecularly defined entities. Seamless access to, aggregation, and integration of these data is essential for data-driven oncology.

Integration of molecular, imaging, and clinical data in cancer research faces challenges due to sparse ontologies and tools for unified analysis. The diverse formats and semantics of these data, stemming from variations in ontologies, vocabularies, and models, create hurdles for cancer data integration. Standards representing phenotypic data are particularly lacking [10]. Ensuring algorithmic use and quality assurance require recorded evidence and provenance. Moreover, there is a lack of tools to facilitate seamless adoption of existing standards during data generation, along with incentives to use them.

The Cancer Mission Europe board report thus recommends investments into comprehensive data integration efforts (recommendation 8) and argues for the creation of the European Cancer Patient Digital Centre (ECPDC).

Cancer Data Harmonisation

Data harmonisation is a major challenge in cancer research. Pooling data from numerous sources to create larger cohorts and strengthen statistical power is only possible when the data are meaningfully connected. Data harmonisation and standards can enable interoperability of multimodal cancer research data, facilitating analysis and knowledge generation for the border research community. Ensuring that the data adhere to the FAIR principles (Findable, Accessible, Interoperable, Reusable) is a key element in facilitating data harmonisation and interoperability [11]. Numerous efforts and initiatives enable the production of FAIR data, both in terms of standards (FAIRSharing), and guidance (FAIR Cookbook, RDM Toolkit).

Application of Artificial Intelligence to cancer research

Building on the FAIR aspects of cancer data, artificial intelligence (AI) stands as an invaluable tool for addressing major open questions in cancer research. For instance, AI has been successfully used to detect cancer biomarkers from mutational or gene expression profiles, enabling diagnosis and early detection based on digital pathology images and predicting treatment responses, among other applications [12].

Although AI promises to be a critical component for the successful development of standardised and automated analysis of patient data in the clinic, its adoption in clinical practice faces numerous obstacles. The relevance of these methods heavily depends on the availability of large-scale well-structured data and well-designed expertly annotated test and training datasets. Variability across different countries, healthcare systems, and legislatures poses significant challenges in creating such coherent datasets. A recent review of AI methods in a clinical epidemiological context demonstrated that the benefits of AI tend to be erased in biased comparisons [13]. Given the potential impact of using an AI model to assist clinical decisions, it is imperative to ensure that such decisions do not reflect biases for or against certain groups or populations or any inherent bias that the provided data may introduce.

Integrating AI into clinical decision-making workflows presents its own challenges. These include the need for standardised validation checks for training datasets, regulatory guidance for medical device generation, and health technology assessment to compare innovative solutions with established practices. While small cohorts' data can be manually curated, AI models typically require large datasets. Therefore, automated, scalable, and reproducible data curation and validation are vital for ensuring successful deployment in medical practice.

Towards a European cancer data federated solution

In light of the aforementioned challenges, it is imperative to establish a cohesive and comprehensive federated European solution for cancer data, embracing the principles of open science and leveraging existing policies and standards to harmonise cancer data management across Europe. To achieve this it would be essential to:

1. *Commit to Open Science principles*, including FAIR data principles for scientific data collection, management and stewardship. This should be one of the foundations of a governance model for cancer data at the European level, in line with the initiative of Cancer Mission Europe to create the European Cancer Patient Digital Center (ECPDC). It will be essential to follow the developments of the proposed legislature for the EHDS since data from clinical settings is increasingly becoming the subject of research.
2. *Foster trust by storing data locally*. Human genetic data and other patient data are considered sensitive personal information and it is essential to harbor trust between

citizens, researchers and clinicians by treating it responsibly and confidentially. Storing data locally would establish a direct and trustful relationship between citizens and the data hosting organisation. Federation between these local implementations could enable a transborder architecture of resources that would allow flexibility in aligning regional and national healthcare organisation and governance across Europe, alleviating legal and citizen trust challenges.

However, federating cancer resources across Europe imposes an additional layer on infrastructure governance. Trust has to be fostered between data hosting organisations themselves. Understanding each institution's national strategic goals and research objectives and how they can be coordinated is essential to maintain and reinforce such relationships, as well as ensure operational efficiency. Such a federated European infrastructure would allow pooling the curation and knowledge-harvesting efforts across Europe, leading to cost scalability.

3. *Draw on existing policies* for sensitive health data sharing across Europe and in alignment with global initiatives such as the GA4GH's Regulatory & Ethics Work Stream intended to support ethical data sharing across jurisdictions. A substantial body of work can already be considered at the European level concerning consent clauses, policy on ethics review, Data Access Committee Review Standards (DACReS) Policy and guidance on Machine Readable Consent.
4. *Build on existing standards* to facilitate the discoverability of information resources and enable the integration of large-scale data collections across many sources. This should include (without being limited to) standards on authorisation and accreditation, ontologies, data representation and APIs. Such unified access to the existing and new data sources is essential to ensure their interoperability and facilitate the adoption of such a platform by future information resources. Moreover, federated cancer analysis portals can be built on top of such APIs, ensuring that the knowledge gained from local data sets becomes seamlessly accessible across the federation of resources.
5. *Enable High-Performace Computing (HPC) on sensitive data*. Federation offers the possibility of leveraging the power of AI on large datasets and implies seamless access to HPC resources, a challenge due to constraints on sensitive data. Possible approaches for secure computations include homomorphic encryption, secure multiparty computation, differential privacy, or federated learning. At a European level, leveraging the EOSC infrastructure for connecting the essential components of cancer data management systems could facilitate interoperability of the federated repositories, along with analytical services and portals for practical use.
6. *Grow competence capacity* by training researchers and clinicians in the digital skills necessary to use the cancer data ecosystem efficiently. Indeed, effectively using tools, platforms and infrastructures that enable effective and efficient data management, provide access to large cancer datasets, and allow execution of interoperable workflows and analytical services is not trivial. The recently published report on "Digital skills for FAIR and Open Science" by the EOSC Executive Board Skills and Training Working Group clearly outlines the different training needs related to the various roles in research. A tailored approach for cancer research would be invaluable in this regard and enable the efficient training of a new generation of cancer scientists.

Keeping these recommendations in mind and leveraging the resources, political drive, and expertise of ESFRIs such as ELIXIR, EATRIS; initiatives such as the European Open Science Cloud (EOSC), 1+MG and the EHDS as well as projects like EOSC4Cancer, GDI and others discussed, would address the challenges in cancer data management, enabling researchers to access and integrate diverse cancer datasets more easily, to achieve a greater understanding of the disease and development more effective treatments. Federated

management and improved interoperability of cancer data would not only advance cancer research but would enable precision cancer medicine to improve outcomes for patients.

Box 1: Cancer Use Cases

Critical use cases for the European cancer data infrastructure include:

1) Combining multimodal data in cancer research

Gathering data from multiple patients and model organisms is crucial to develop and validate actionable prognostic and predictive models for precision treatment decisions. Harmonising data from multicenter studies requires significant effort to establish and implement standards across all sites using containerised analysis workflows (Figure 3, left).

All data needs semantic annotation (e.g., Gene Ontology, SNOMED CT), including multi-modality data such as gene expression, genetic variation and clinical imaging. Minimising batch variation across sites requires careful data normalisation. Identifiers should be removed according to agreed procedures to prevent reidentification, and all datasets could then be securely combined for downstream analyses. Molecular and pathway signatures for tumor subtypes can be refined and extended from the unified and normalised omics datasets. Statistical evaluation can associate these signatures (e.g., an immune signature) with phenotypes (e.g., HPV positivity) and clinical outcomes. Finally, one can compare the performance of a deep neural network trained on the collected data with that of networks trained on public datasets. Model results must be assessed against the standard of care (SoC) in a defined validation dataset to determine applications for patient stratification and personalised treatment.

2) From research to molecular-guided clinical trials

Efficient patient-treatment matching in clinical trials relies on a research-enabled understanding of molecular and pathway signatures. Cancer patients' clinical, diagnostic, and treatment data, along with genetic profiling information, are stored in dedicated Electronic Health Record (EHRs) or linked systems (Figure 3, right).

Since the number of similar cases in a single hospital is limited due to the diversity of tumour characteristics, integrating the hospital's system with international datasets from collaborating hospitals can harness the power of a larger combined cohort. Larger datasets enable the utilisation of research-defined criteria, such as molecular and pathway signatures, to identify similar patient subsets for hypothesis testing, biomarker discovery, and validation. Data-sharing among hospitals (Figure 3, center) also offers possibilities that may be otherwise unavailable due to resource constraints. For instance, in combined basket/umbrella clinical trials with multiple arms, the decision to open a new arm depends on the likelihood of recruiting enough patients with specific molecular profiles within a relevant time frame. Querying aggregated data helps determine the feasibility of opening a new arm based on patient inclusion numbers, enabling targeted treatment options based on molecular data for newly enrolled patients.

3) Personalised treatment based on a patient journey

The patient journey encompasses the entire process that a patient goes through from initial health issue detection, diagnosis by healthcare professionals, disease awareness, to the treatment aiming for a cure or disease management. From the moment the patient is admitted to the hospital, different medical tests and procedures are performed for clinical decision-making, generating multiple heterogeneous data at several time points (e.g., diagnosis and follow-ups) included in the patient's EHR (Figure 3, right).

The importance of the clinical decision support system integrating molecular and non-molecular data from large-scale cohorts and assessed by AI predictive software cannot be overstated. Such a decision system is a time-saving aid for selecting appropriate diagnostic tests, making informed treatment decisions, facilitating the interpretation of the response to treatment, and orienting the patient towards relevant clinical trials. Thus, to ensure their pertinence, clinical decision support systems have to contain all the relevant patient data, such as omics data (e.g., somatic mutations), digital pathology (histology, imaging) and other clinical data (e.g., treatment history), a well as links to data from large aggregated cohorts, databases containing actionable mutation list, approved drugs, known resistance mutations, and even ongoing clinical trials.

References

[1] Europe's Beating Cancer Plan, Communication from the commission to the European Parliament and the Council, COM (2021)44

[2] European Commission, Directorate-General for Research and Innovation, Pita Barros, P., Beets-Tan, R., Chomienne, C., et al., Conquering cancer: mission possible, Publications Office, 2020, https://data.europa.eu/doi/10.2777/045403

[3] Eggermont, A.M., Apolone, G., Baumann, M., Caldas, C., Celis, J.E., de Lorenzo, F., Ernberg, I., Ringborg, U., Rowell, J., Tabernero, J. and Voest, E., 2019. Cancer Core Europe: a translational research infrastructure for a European mission on cancer. Molecular oncology, 13(3), pp.521-527.

[4] O'Doherty, K.C., Shabani, M., Dove, E.S., Bentzen, H.B., Borry, P., Burgess, M.M., Chalmers, D., De Vries, J., Eckstein, L., Fullerton, S.M. and Juengst, E., 2021. Toward better governance of human genomic data. Nature genetics, 53(1), pp. 2-8.

[5] Freeberg, M.A., Fromont, L.A., D'Altri, T., Foix Romero, A., Izquierdo Ciges, J., Jene, A., et al (2022) The European Genome-phenome Archive in 2021. 50(D1), D980–D987.

[6] Rehm, H.L., Page, A.J., Smith, L., Adams, J.B., Alterovitz, G., Babb, L.J., Barkley, M.P., Baudis, M., Beauvais, M.J., Beck, T. and Beckmann, J.S., 2021. GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell genomics, 1(2), p.100029.

[7] Eiss, R. (2020). Confusion over Europe's data-protection law is stalling scientific progress. Nature, 584(7822), 498-499.

[8] The PHG foundation (2020). The GDPR and genomic data. https://www.phgfoundation.org/report/the-gdpr-and-genomic-data

[9] Jiang, P., Sinha, S., Aldape, K., Hannenhalli, S., Sahinalp, C., & Ruppin, E. (2022). Big data in basic and translational cancer research. Nature Reviews Cancer, 1-15.

[10] Deans, A.R., Lewis, S.E. Huala, E., Anzaldo, S.S. Ashburner, M., Balhoff J.P., et al. (2015). Finding our way through phenotypes. PLoS Biology, 13(1), e1002033.

[11] Kush, R. D., Warzel, D., Kush, M. A., Sherman, A., Navarro, E. A., Fitzmartin, R., et al. (2020). FAIR data sharing: the roles of common data elements and harmonization. Journal of Biomedical Informatics, 107, 103421.

[12] Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 20(5), e262-e273.

[13] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Figure Legends

Figure 1. Health "data journey" between clinical decision-making and translational research and current challenges. Local management and controlled access to health-derived multi-source data can fuel global scientific discoveries to help bridge the gap between translational research, informed clinical decision-making and practical applications in health innovation.

Figure 2. ELIXIR's joint effort with other human data initiatives. Through collaborative actions, ELIXIR aims to harness key partners' and other EU initiatives' collective expertise and resources, fostering a holistic approach to gathering, managing, and analysing clinical and biomedical data. By joining forces with other initiatives focused on human data, ELIXIR strives to enhance data infrastructure, accessibility, quality, and interoperability, ultimately accelerating breakthroughs in biomedical research and promoting data-informed decision-making.

Figure 3. Data flow and sharing: from cancer research to personalised precision medicine and secondary uses. Federated local repositories located in healthcare institutions would allow the storage of cancer patients' multi-modal data, allowing controlled remote access (via web portals or APIs) for secondary use. This high-volume data scenario enables the development of personalised precision medicine in cancer by facilitating the effective training, refinement and validation of algorithms for classification, diagnosis and prediction of patient's evolution and prognosis, as well as the improvement of tools for automatic interpretation of multi-omic tumour data, clinical images, clinical decision-making support.

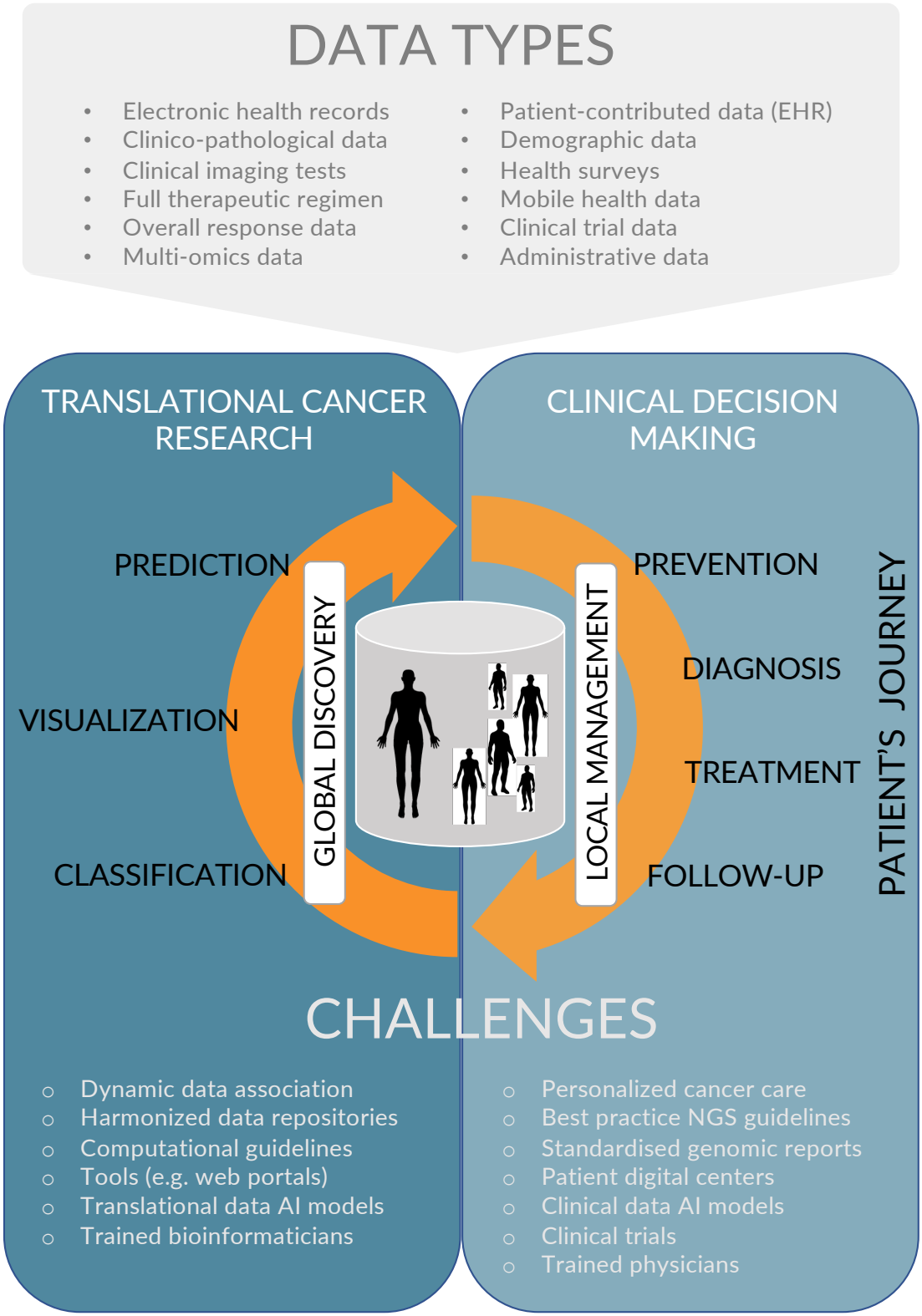

Figure 1 Health "data journey" between clinical decision-making and translational research and current challenges.

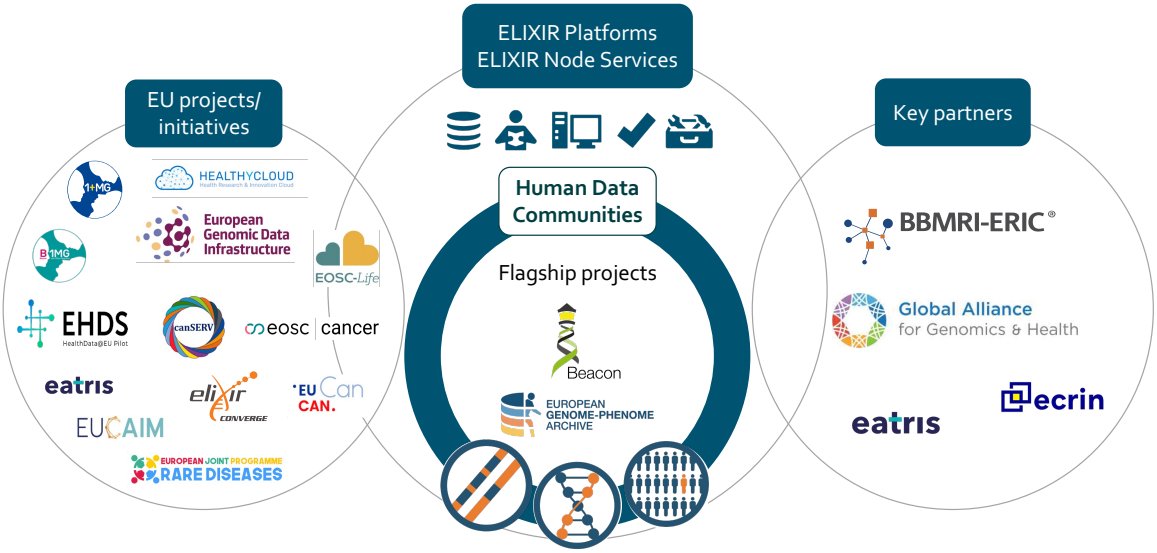

Figure 2 ELIXIR's joint effort with other human data initiatives.

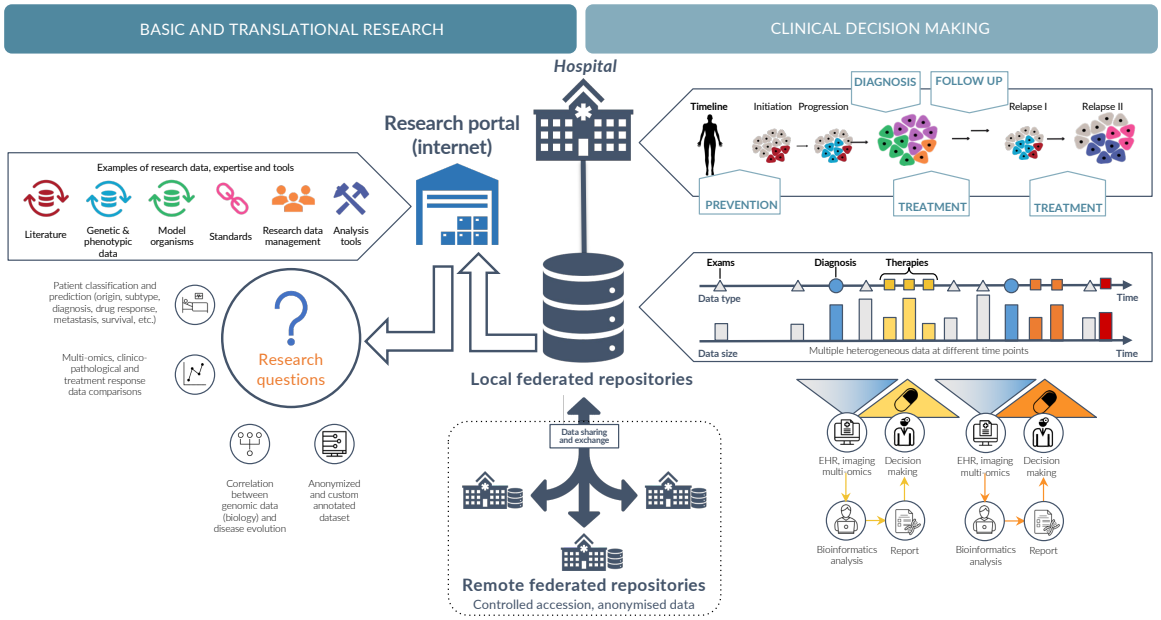

Figure 3 Data flow and sharing: from cancer research to personalised precision medicine and secondary uses