RÉSEAU EUROPÉEN **DE RECHERCHE TRANSLATIONNELLE ET D'INNOVATION EN ONCOLOGIE**

ONCONET SUDOE

# GUIDE FOR DATA MODELIZATION, INTEGRATION AND EXCHANGE IN ONCOLOGY

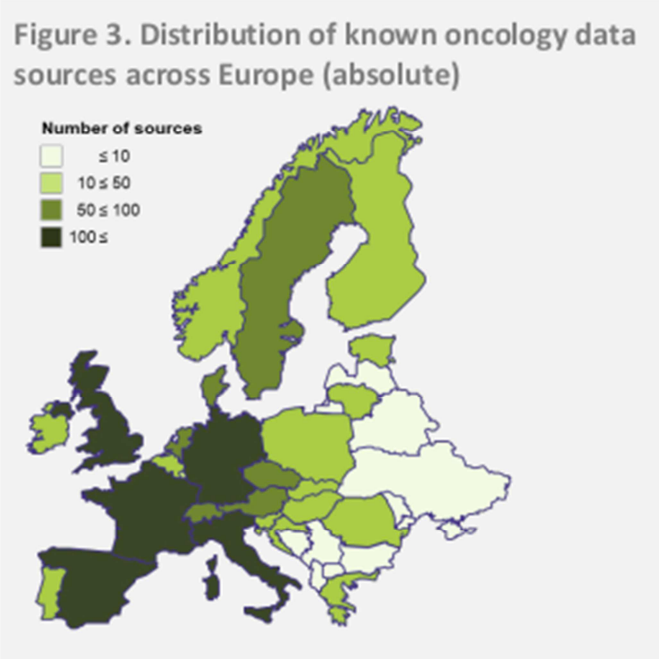
# Table of contents:

*With the advent of new technologies, health data is increasingly available from a wide range of sources. This explosion of information offers new opportunities but also highlights old challenges, such as the fragmentation of the European data landscape, quality and methodological limitations, and concerns around data privacy and security, to name but a few. Given the potential of data to improve population outcomes and health system sustainability, these challenges must be overcome.*

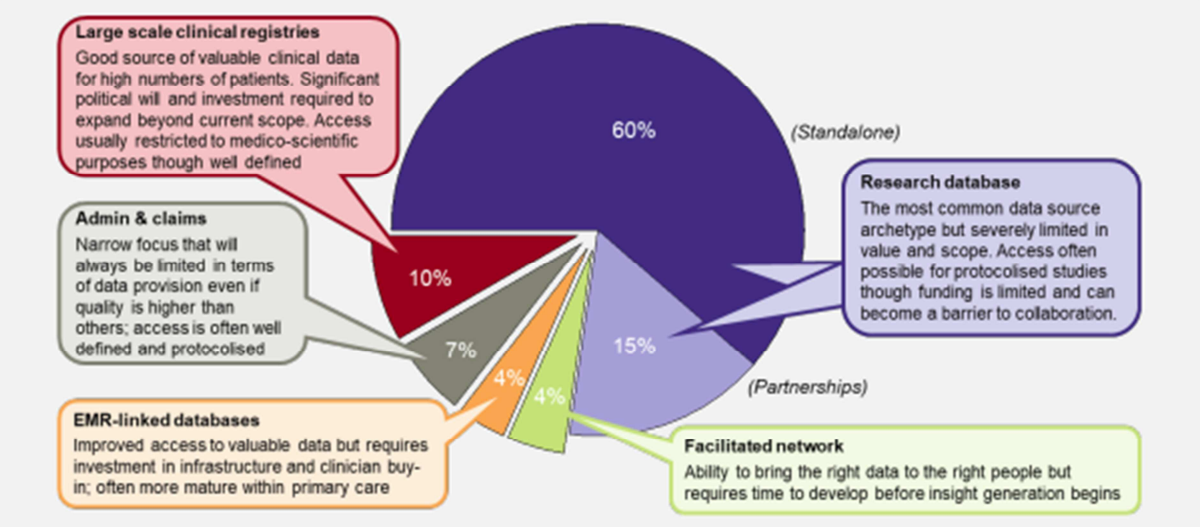
*This is particularly critical in oncology, which has seen unprecedented innovation in recent years: treatment paradigms are shifting from tumour- to mutation- and biomarker-based, gene editing is now within the realm of the possible, but data sources and standards are struggling to keep pace. Moreover, certain aspects of oncology – such as the reliance on genetic information and biomarkers, narrow patient populations, and the growing promise of combination therapy – add layers to the already complex European data landscape.*

# Oncology data sources in Europe

European healthcare systems reflect centuries of political decisions, economic challenges and cultural mindsets to name but a few. There are national-led systems (National Health Insurance or Sécurité Sociale in France, National Health Service (NHS) or Servico Nacional de Saude (SNS) in Portugal) and regional ones (Spain). There are tax-funded systems (Portugal, Spain) and insurance-based models (France).




The European landscape is characterized by a multiplicity of data sources. There are over 1,100 oncology data sources across Europe. Almost 75% of oncology data sources in Europe are either standalone or partnership academic registries, while the remainder is made up of large-scale clinical registries, administrative data and claims, facilitated networks, and electronic medical records. About 80% of these sources cover cancer data.



Data integration related initiatives in Europe

We have identified more than 40 data initiatives surveyed across Europe, 35% have a focus on improving collation, and 25% have a focus on improving access. 38% have more than one objective, highlighting the need for comprehensive approaches to improve the health data landscape. Public or grant support, including European Commission financing and/or governance, is a source for 70% of initiatives. Although a therapy area focus is frequent, 30% of initiatives are disease-agnostic, 25% address cancer-specific issues overall, 23% cover more than one tumour type and 23% focus exclusively on one cancer type.

These initiatives have brought much improvement to the European oncology data landscape, e.g. by promoting the use of the **Observational Medical Outcomes Partnership (OMOP) model** – a common data model enabling the comparison of data collected in different formats.

| Improve Access | Improve Collation | Standardise Data | Collect New Data Types |
|---|---|---|---|
| *Aims to improve access to existing datasets or allow their interrogation* | *Aims to incorporate existing datasets into a central repository* | *Aims to standardise the ways in which data is collected so that datasets re comparable* | *Aims to collect data that does not yet exist, often via novel approaches* |
| • BD4BO<br>• CODE<br>• GOBDA<br>• HemoBase<br>• IMI Harmony<br>• INSITE<br>• PHEDRA<br>• POI<br>• Simulacrum | • Cancer Core Europe<br>• ECIBC<br>• ECIS<br>• EUROCARE<br>• HMRN<br>• ENCR<br>• EUCAN<br>• EUSOMA<br>• Greater Manchester Cancer<br>• IMI Protect<br>• Innovative Pricing Solutions<br>• I-O Optimise<br>• REAL Oncology<br>• Sarcoma BCB | • EHDN<br>• GA4GH<br>• GEKID<br>• FRANCIM<br>• Health Informatics Collaborative<br>• ICHOM<br>• OMOP Oncology | • 100,000 Genomes Project<br>• AURORA<br>• EUROSTAT<br>• CRISP<br>• IRONMAN<br>• OWise<br>• My Clinical Outcomes<br>• SCAN-B<br>• Universal Cancer Databank<br>• WEB-RADR |

# Current challenges with oncology health data in Europe

In this chapter we will see and discuss the major challenges linked to health data. For each challenge an analysis and some clues on how to address it (good practices) will be provided. These challenges could be summarized as follow:

| | |
|---|---|
| **Data** – requiring broader, deeper & interoperable datasets | • Most current data sources do not collect all the relevant information that is useful in oncology (e.g. biomarkers, ECOG scores, surrogate endpoints, etc.)<br>• Much of the data used to manage cancer is unstructured and coded in different languages<br>• Data quality varies across datasets, due in part to insufficient quality control mechanisms |
| **Structure** – needing a stable, open and supportive environment for data | • Few cancer plans and country policies explicitly target health data<br>• Member states have the ability to legislate locally for health, genetic and biometric data<br>• Funding for health data tends to be short-term and come with individual interests<br>• Linkage is key to enrich decision-making but remains limited for legal, societal and technical reasons |
| **Process** – progressively scaling up to world-class, transparent processes | • Access requirements differ widely across datasets<br>• Patient consent is not optimised and can both delay and limit the availability of data<br>• The timeliness of data availability and access is a significant problem<br>• Significant resources must be expanded to protect patient data and privacy |
| **Technology** – enabling solutions that were previously difficult | • The lack of interoperability between systems limits the ability to link different sources of health data<br>• Software and platforms are rarely user-friendly, limiting the ability to collect enough data<br>• Existing technologies may be outdated or likely to become so given new processing requirements |
| **People** – building skills & mindsets for involvement & sharing | • There is a lack of awareness and misconceptions undermine the full potential of health data<br>• Patient concerns around their privacy remain strong<br>• Healthcare providers remain concerned about the use and sharing of their patients' data<br>• There is a lack of qualified individuals to undertake the increasingly complex task of collecting data<br>• Vested interests and stakeholders' own agendas hinder collaboration in private and public settings |

**Data – requiring broader, deeper and interoperable datasets**

Cancer is a complex disease, in which an improved understanding of patients' genotype and the impact on disease progression have been playing a growing role. Where only a few years ago treatment was based on the location (e.g. breast, lung, colon) and type of the tumour (e.g. sarcoma, leukaemia), drug regimens are now prescribed based on genetic mutations.

- Most current data sources – established years ago – do not collect patients' DNA, nor the additional information that is useful in oncology such as biomarkers, Eastern Cooperative Oncology Group (ECOG) performance scores and surrogate endpoints (e.g. PFS).
- Recent years have seen some initiatives, such as the UK's 100,000 Genomes Project or US-based Flatiron Health, develop to collect this information for various cancer types, but these initiatives are in early stages.

Much of the valuable information used in cancer management is unstructured – it is not stored as coded inputs, but collected as notes, voice recordings, scans, histological stains, etc. that are virtually impossible to compare systematically across datasets using conventional technologies.

- These use different languages across countries and different coding standards (e.g. DICOM, WADO, HL7 are used across Europe), if at all.

- Machine learning could help bridge this gap and there are some efforts to align coding standards – for example, the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) sets standards for genomic biomarkers as part of topic E15.
- The use of ETL (Extract Transform Load) architectures could help establishing the links between several information sources but could lead to weak the information accuracy by blending frames of reference.

Like unstructured data, structured data can be coded in different ways, limiting the ability to compare datasets across sources and countries.

- France, Germany and Austria use ICD-10 for diagnosis, Denmark and Finland also use ICPC and ICPC2, and Belgium and the UK use ICD-10 but also SNOMED.
- The World Health Organisation (WHO) is publishing guidelines to encourage wider adoption of the International Classification of Disease (ICD) coding framework.
- Federated querying can pool comparable available data in different formats, and open, cloud application programming interfaces (API) like that launched by Google in 2018 can help manage medical datasets in multiple standards.

Further limitations of data collection protocols further limit the relevance of data collected, though there are ongoing efforts to overcome this.

- The exact data differs across and within countries, with only some Member states (e.g. Belgium, Poland, Spain) defining minimum datasets for their public databases.
- In some cases, the data collected lacks the requisite granularity: for example, registries tend to collect only first-line treatment.
- In the US, the FDA developed a guideline on biomarkers in 2005, and its ICH E15 version defines coding standards for biomarkers.

Data quality varies across datasets, due to incomplete coverage (i.e. when data is not or partially entered for patients) or inaccuracies (i.e. when data is erroneously entered).

- Current quality control mechanisms are not sufficient to address this: there is often unclear responsibility for quality assurance between the EU and its Member states, and many countries do not have specific legislation on data quality (e.g. Bulgaria, Estonia, Greece) or mandatory quality audits of EHRs (e.g. France, Germany, Netherlands, Sweden).
- Some countries have recognized this as an issue, with the NHS England recommending a clinical coding audit to take place every 12 months, and countries like Iceland and Estonia implementing EHR audits.
- In Belgium, the government has incentivized healthcare providers with almost €900 per head to subscribe to an EHR system with decision aids and categories to help enter the right data.

**Structure – needing a stable, open and supportive environment for data**

The lack of clear political will and direction across European, national and regional health structures has led to the development of numerous, uncoordinated sources – these emerge to fit individual, ad hoc needs, often lacking clear standards and duplicating work done elsewhere (due to the lack of visibility of many ongoing efforts).

- There are no European cancer plans, and most national cancer plans barely touch upon health data, though some – like the French Plan Cancer 2014-2019 – do.

- In regional health systems, individual cancer plans and data policies vary widely, if they are even defined, leading to disparities in health data use across regions: for example, Lombardy is leading the way in Italy, and Catalunya and Andalusia in Spain)
- Even with the GDPR aiming to harmonise data privacy laws across Europe, Member states retain the ability to legislate locally for health, genetic and biometric data, which is unlikely to lead to European alignment.
- EU-wide efforts, such as the European Patient Smart Open Services (epSOS, launched from 2008-2014) or Cancon (2014-2017, co-funded by the EU Health Programme to improve the quality of cancer control, including cancer data, across Europe) have been launched; some can be short-lived and add to an already complex landscape, while others like epSOS become more widespread and underpin new best practice .

Health data infrastructure is not available across all settings of care and European countries, and where it is, it often needs to be updated, requiring significant resources to update and maintain.

- Funding continues to be an issue for many data sources, initiatives and innovations, which struggle to attract and retain neutral, long-term funding – 14% of healthcare providers see funding as the main eHealth challenge they face, reaching ~30% in Ireland, Austria and the UK.
- This is primarily due to the short-term nature of investment, with funds typically running out after a set number of years, uncertainty around which entity then steps in, and possible loss of the longitudinal value of data if the initiative ends: for example, funding for EUROCARE stopped in 2018 and the researchers are writing up the final manuscript.
- Investment often comes from several sources, each with their own objectives – this is the case for EHR funding for the NHS England, which is provided by the Integrated Digital Care Fund, Nursing Technology Fund, the NHS Innovation Scheme and Vanguard sites.
- Processes to obtain funding may be unclear and complex, requiring time or even a government partner to understand how granting agencies evaluate proposals
- Some centrally-funded initiatives cannot apply for external funding.

At a more granular level, the data infrastructure is such that linkage is essential to enrich insights gathered from data but remains limited for societal and technical reasons. Indeed, many are concerned about the ability to identify patients by connecting sufficient information about them, possibly leading to stigmatisation by peers or profiling by insurers. As a result, many countries (e.g. Germany, Portugal, and Norway) report limited linkage of databases.

- This can be due in part to the lack of single identifying numbers to link relevant data, or to their limited scope: in France, identifying numbers used by hospitals can vary across hospitals and are different from those used for medical insurance.
- In addition, in many countries linkage must be allowed on a case-by-case basis by a dedicated authority (e.g. the Privacy Commission in Belgium), or by national or local legislation (e.g. the law in the state of Bremen in Germany authorises linkage).
- Nonetheless, many countries successfully use single identifying numbers, providing comprehensive insight into patients' health that extends beyond traditional data: the Swedish 'personnummer' is unique in being used for purposes as diverse as tax, social welfare, health care, living conditions and education.
- In 2014, the eIDAS regulation was approved by EU co-legislators to ensure that people and businesses can use their own national electronic identification schemes

to access public services in countries where EUid are available, and creates an European internal market for time stamps and other means of authentication.

**Process – progressively scaling up to world-class, transparent processes**

Processes to collect, use and share data often require clear justification to protect patients. In the context of the GDPR, this rationale has been more clearly explored but can still pose challenges.

- The use of RWD without patients' consent is justified on medical care and public health grounds, but how this is interpreted locally and which data users may qualify for these uses will vary, possibly leading to fines.
- Secondary use of data (i.e. use of data not explicitly collected for this use) is typically acceptable in the context of scientific research, but is currently regulated locally: some countries may accept other objectives for secondary use, such as Belgium where data can be used for historical and statistical analyses, but these are not the norm.
- Many data initiatives feel that they sufficiently cover these grounds, and should therefore be able to continue using their data, but only time will tell whether that is true and the data remains available.
- As part of its Public Sector Information proposal, the European Commission aimed to increase the availability of data by bringing new types of public and publicly funded data into the scope of the Directive, and provide that data already available in "open access" research data repositories should be re-usable for commercial and non-commercial purposes.

Processes surrounding health data are often unclear, time-consuming and restrictive.

- Access requirements differ widely across datasets, limiting stakeholders' ability to readily consult multiple sources by re-using similar materials and applications, and various bodies may be required to approve access – in Germany, 16 regional data protection agencies review data access protocols.
- Although most European databases are accessible to academics upon request, some data sources only grant access via third parties or offer limited access to private entities: in France, the Programme de Médicalisation des Systèmes d'Information (PMSI) hospital claims databases can only be accessed via a neutral third-party provider.
- The more stakeholders are involved in an initiative, the more cumbersome the process; different governing members may also be more conservative than others within a single initiative.
- In some cases, requirements and processes are so complex that initiatives had to stop using data due to changes in third-party access requirements.

Patient consent remains at the core of data collection, but it is not optimised and can both delay and limit the availability of data.

- Only 13 of the 28 EU countries have specific rules regulating patients' consent for EHRs, while frameworks and best practice tend to remain local.
- Consent processes differ widely, are often unclear for patients and can be quite complex: consent forms for research can range from three to 30 pages, with an average readability suitable for a college graduate .
- Opt-in consent management solutions, user-friendly videos and other tools have been used to facilitate the consent process, allowing patients to have a better view of the data they offer and how this is used for different applications; in the Nordics

and Belgium, non-sensitive identifiable personal data can be made available to researchers without prior consent for research purposes, accelerating access.
- The GDPR sets out to delineate stronger and clearer conditions for consent, but if not done properly, this could also increase the amount of information that patients are required to consider, understand and agree to.

Across European countries and data sources, the timeliness of data availability and access is a significant problem – it can take up to four years to obtain data, even though cancer evolves on a daily basis and time is of the essence.

- Contract signing, scientific and ethical approval are often slow, requiring multiple parties' review.
- Depending on databases' and studies' set-up, scope and software used, data can take months to be collected and cleaned up, let alone accessed by those having followed the right process: for example, one national cancer registry's ethics and access agreements currently take more than six months, with data that is more than 18 months old due to requirements for consolidation and learning.
- To enable optimal decision-making in oncology, timely information is critical – to this effect, initiatives like the Collaboration for Oncology Data in Europe (CODE) aim to provide real-time data for treatment decision-making and new payment models.

Beyond ethics, consent, analysis and access processes, significant resources must be expanded to protect patient data and privacy.

- Patient data can be de-identified, but this is not infallible; full anonymization may be necessary but can be challenging, requiring multi-stage de-identification with clear governance and controls approved by relevant authorities.
- Data aggregation helps overcome some privacy concerns but may not provide the granularity of information required for some decisions.
- If accepted by decision-makers, simulated data based on real patient characteristics could provide a way through and inform data-querying models and decision-making, though this would be limited in scope to hypothesis generation and methods validation in the near future – Simulacrum, a collaboration between Public Health England, Health Data Insights, IQVIA and AstraZeneca, provides simulated data modelled from the Cancer Analysis System.
- New technologies can also foster better data privacy, but remain in their early stages: for example, blockchain (a list of data blocks secured by complex codes and accessed via a transparent ledger) can support a clear audit trail across a secure platform with decentralised ownership; a number of start-ups are already operating in this space: in 2017, Estonia's eHealth Authority signed a deal with Guardtime to secure the health records of its citizens.

**Technology – enabling solutions that were previously difficult**

Technology provides unprecedented opportunities to generate more, higher-quality health data, and share it with all the relevant parties, but barriers remain well beyond oncology. The lack of interoperability between systems limits the ability to link different sources of health data.

- In the UK, there are more than 100 commercial suppliers of EHR software, let alone for other sources for health data.
- In France and Spain (despite of the NHS), most hospitals develop their own, fit-for-purpose software with limited intent to connect with other databases.
- Across the EU, only 13 countries have set up specific rules on interoperability (e.g. Austria, Belgium), and only six for cross-border interoperability (e.g. Spain).
- In 2018, the European Commission announced the development of technical specifications for a European EHR exchange format, to enable the European data space.
- Numerous initiatives such as O-Wise and the Haematological Malignancy Research Network (HMRN) have been established to enable linkage, but these are still not the norm.

Software and platforms for health data are rarely user-friendly, limiting the ability to collect sufficient high-quality data.

- Out of 38 papers on EMR implementation, seven listed ease of use as a main barrier.
- Interviews with oncologists in France, Italy and Spain emphasized the complexity, low user experience, and high requirement for manual processing across their hospitals' EMR systems.
- To bypass this, successful data sources and initiatives often employ dedicated technicians for data entry.
- Recent years have also seen improvements in the quality ratings for US-based EMR interfaces and visual appeal, suggesting a democratisation of these software and a growing focus on the users' experience; this could potentially increase adoption in Europe, as well.

Existing technologies may be outdated or likely to become so in view of growing analytical and processing requirements.

- Some adjustments will be required – between 100 million to two billion human genomes could be sequenced by 2025, requiring 2-40 exabytes of storage capacity and processing that is six orders of magnitude faster than is possible today.
- Technology is already stepping in to meet this gap: for example, cloud computing solutions can be used for large-scale analysis and storage of health data, enabling continuous coordination of patient-care and seamless integration with health systems; machine learning can help automate part of the data entry process.
- Awareness, understanding and mastery of these technologies currently remains limited to more advanced IT and digital companies, highlighting the need for increased availability, remuneration and training in these skills.
- Updating infrastructure is costly but can be preferable to a complete system overhaul that could jeopardise hospitals', cancer centres and general practices' continuous flow of data – the balance between 'building from scratch' and incrementally upgrading will depend on the systems, capabilities and requirements of each data stakeholder.

**People – building skills and mindsets for involvement and sharing**

There is a clear lack of qualified people to undertake the increasingly complex and comprehensive task of collecting and analysing data.

- Although some initiatives provide specific training for employees, the lack of data scientists and professionals qualified in new technologies (e.g. analytics, machine

learning) is the top issue for 7% of European healthcare providers, and is strongest in the public sector.

- Many healthcare professionals have limited digital literacy or training in data collection, in addition to having numerous other responsibilities – in the UK, poor staff engagement and training led to the Cambridge University Hospital Trust reverting to paper records after an attempt to roll out 2.1 million EMRs in 2014.
- More simply, limited manpower can also be a challenge: data collection and analysis are resource-intensive, and availability of staff can limit data initiatives' ability to scale up.
- Dedicated courses and degrees on analytics, data sciences and digital health skills are emerging across most European universities, with companies like IBM setting up partnerships with universities and funding for more data science training.

As health data becomes more of a business, vested interests and stakeholders' own agendas may sometimes hinder collaboration across both private and public settings. Data sources and initiatives, which also spent a lot of time and effort collecting and cleaning up data, can also be reluctant to share it.

- Private pharmaceutical, medical device, biotech and/or technology companies protect their commercial interests, but health insurers – including state-owned ones like Germany's statutory health insurances – and other publicly-funded entities also limit the potential use of their data.
- Although the GetReal melanoma case study was funded by EFPIA, EMA, the UK's National Institute for Health and Care Excellence (NICE) and the Dutch National Health Care Institute (ZIN), some participating registries restricted access to enable their PhD students to publish their theses with that information .
- In the Netherlands, the Dutch Upper GI Cancer Group has established a process to enable sharing with any who ask for data: a committee reviews applications to access their data, whose members can oppose access, but this rarely happens and the data is readily shared.
- Recognising the value of this data, pharmaceutical companies are increasingly partnering with and acquiring data sources, as demonstrated by Roche's acquisition of FlatironHealth, an oncology-focused EMR company.

Beyond commercial interests, healthcare providers remain concerned about the use and sharing of their patients' data, and may not be aware of ongoing data initiatives or mechanisms in place to protect patient confidentiality.

- In the UK, the NHS's care.data scheme – designed to unify patients' care across general practices and hospitals into one central database – was postponed and subsequently cancelled due to physicians' opposition to privacy and consent issues.
- In France, the Dossier Médical Partagé – an initiative to ensure every French patient has a medical record – reached 400 thousand records within two years, well below the objective of 500 thousand records within one year; this was primarily due to the lack of awareness or campaigns geared towards physicians.
- Recognizing lack of health professional engagement as a barrier, several governments are partnering with trusted entities and collaborating more closely with physicians to foster better buy-in in data initiatives: the Belgian government is collaborating with Custodix, a trusted third party EMR vendor with a strong reputation for data hosting and transfer, thereby inspiring trust and reducing resistance towards collection of health data.

Patient concerns around their privacy remain strong, particularly given recent scandals and data breaches.

- Only 38% of EU patients believe that healthcare providers offer effective data security, and many fear that their data could be used for profiling by insurers (e.g. to increase premiums).
- As a result, numerous efforts to collect patient data meet continued opposition or have failed: the introduction of health e-cards was delayed over 15 years in Germany.
- In the Netherlands, a publicly-funded initiative to that aimed to build a national EHR system to facilitate patient-level information exchange between care providers failed due to opposition from patient groups around data privacy issues during information exchange.

Until the public and patients specifically can fully envision the benefits of collecting and using health data, and be assured that it will be well protected, the health data landscape will continue to have mixed perceptions within what must be its most important supporter group. European citizens and patients are also taking a growing role in the generation and utilisation of health data, which provides unique opportunities to understand real-life health events, choices and perspectives.

- Lack of awareness and misconceptions undermine the full potential of health data: even in healthcare, many individuals cannot readily point to the benefits of health data.
- Patients are rightfully concerned by the use of their data where there is no clear public benefit and solely commercial motivation, but 60% of UK patients would rather grant access to their data to commercial entities than miss out on benefits deriving from their innovation.
- Recent research and campaigns such as #datasaveslives are attempting to fill that gap, but given the omnipresence of data across healthcare systems, much more could be done to quantify its impact on patients and institutions.

# Data modelling

Medical data models describe data structures of information systems in medicine. In routine healthcare a disease-specific data model is needed to address all relevant patient attributes. Approximately 400 data elements are needed per diagnosis in routine healthcare, corresponding to more than 5 million data elements. The Systematized Nomenclature of Medicine Clinical Terms (SNOMEDCT) contains >300000 non-synonymous concepts, i.e. there are at least 300 000 options for a data element.

Due to the complexity of medical terminology, the overall number of medical data models is very high and the vast majority of these models are not available to the scientific community. A recent initiative, the Portal of Medical Data Models (MDM, https://medical-data-models.org) has been created to foster sharing of medical data models. MDM is a registered European information infrastructure. It provides a multilingual platform for exchange and discussion of data models in medicine, both for medical research and healthcare. MDM contains 4387 current versions of data models (in total 10 963 versions). 2475 of these models belong to oncology trials. The most common keyword (n 1⁄4 3826) is ‘Clinical Trial’; most frequent diseases are breast cancer, leukemia, lung and colorectal neoplasms.

The conceptual data modeling process involves the following:

1. Normalize domain concept entities and research object entities to allow maximal data development flexibility for both;
2. Use a many-to-many relationship data structure to associate these two types of entities for fact data collection, and utilize temporal and spatial stamps to annotate fact data;
3. Ensure individualized data integrity and continuity by enforcing an “is a part of” (one-to-many) data relationship between a person and biomaterials derived from the person.
4. Separate data values from data structures to assure that data can be queried across domains within the warehouse and retrieved from the warehouse effectively.

# Research Data Alliance recommendations: The FAIR Guiding Principles for scientific data management and stewardship

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles.

These Principles define characteristics that contemporary data resources, tools, vocabularies and infrastructures should exhibit to assist discovery and reuse by third-parties. By minimally defining each guiding principle, the barrier-to-entry for data producers, publishers and stewards who wish to make their data holdings FAIR is purposely maintained as low as possible. The Principles may be adhered to in any combination and incrementally, as data providers' publishing environments evolve to increasing degrees of 'FAIRness'. Moreover, the modularity of the Principles, and their distinction between data and metadata, explicitly support a wide range of special circumstances.

**Box 2: The FAIR Guiding Principles**
**To be Findable:**
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
**To be Accessible:**
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
**To be Interoperable:**
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
**To be Reusable:**
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

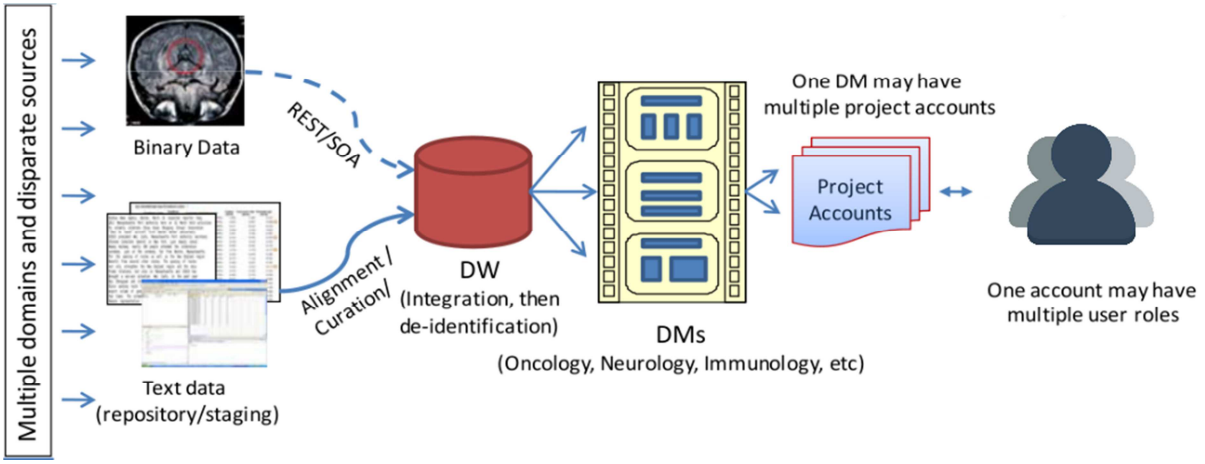
These high-level FAIR Guiding Principles precede implementation choices, and do not suggest any specific technology, standard, or implementation-solution; moreover, the Principles are not, themselves, a standard or a specification. They act as a guide to data publishers and stewards to assist them in evaluating whether their particular implementation choices are rendering their digital research artefacts Findable, Accessible, Interoperable,

and Reusable. We anticipate that these high level principles will enable a broad range of integrative and exploratory behaviours, based on a wide range of technology choices and implementations.
The FAIR principles are already being applied in a range of health-related projects. The FAIR4HEALTH (https://www.fair4health.eu/) project, for example, proposes to apply the FAIR approach in order to accelerate the sharing of data from public research.

# Data Integration

Information systems are a key success factor for medical research and healthcare. Currently, most of these systems apply heterogeneous and proprietary data models, which impede data exchange and integrated data analysis for scientific purposes. Data integration is a constant challenge in translational science. In the past decade, several data integration regimes, including federated database strategies, workflow approaches, semantic web, and warehousing methods, have been tested in the biomedical informatics community. The strengths and limitations of these approaches have been carefully reviewed and contrasted with our local cancer translational research community has further specified their informatics demands, which can be categorized as follows:

1. Researchers want to be able to search and retrieve semantically and descriptively consistent data across domains and longitudinally, and use these data for quantifiable analysis with little or no additional effort for data manipulation and cleansing.
2. Bio-specimen data need to be annotated with available clinical and translational research data.
3. Molecular research (e.g., genotyping) and phenotypic (clinical) records should be interlinked at the level of individuals if they are derived from the same persons.
4. Researchers demand to protect their data privacy for ongoing research but also want to be able to share these data with collaborators.
5. Researchers hope they can curate and annotate integrated data, and eventually develop an evidence-based knowledge base for all cancers.



# Data Anonymization

Anonymization, sometimes also called *de-identification*, is a critical piece of the healthcare puzzle: it permits the sharing of data for secondary purposes. The purpose of this section is to walk you through practical methods to produce anonymized data sets in a variety of contexts.
New algorithms using holomorphic coding could be used to provide with anonymized data which ensures the preservation of a set of properties. These algorithms allows to maintain data consistency while erasing means to individualize contents.
https://hal.inria.fr/hal-01285073/file/978-3-662-43936-4_BookFrontmatter.pdf

https://www.oreilly.com/library/view/anonymizing-health-data/9781449363062/ch01.html
https://es.slideshare.net/kelemam/ahd-oreilly-webcastpart12v3
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5504813/

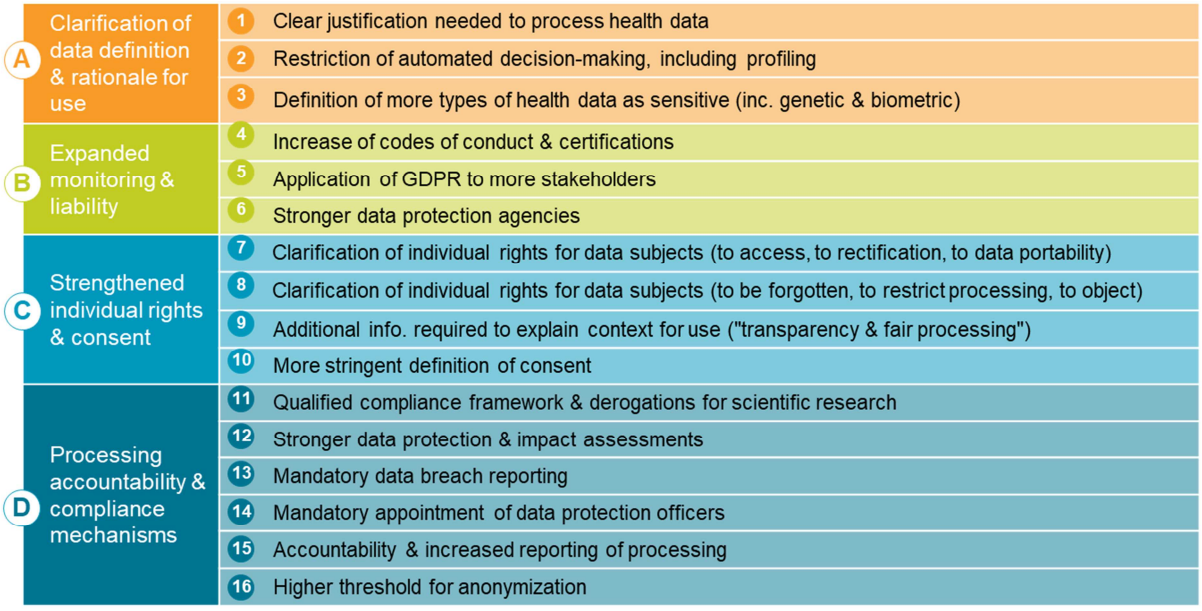
## GDPR

The GDPR is the new legal framework in the EU that aims to:

- Harmonise data privacy laws across Europe
- Protect and empower all EU citizens
- Reshape the way organisations across the region approach data privacy

The regulation came into effect on 24th May 2016, but took full effect on May 25th 2018, replacing the Data Protection Directive 95/46/EC. It establishes minimum mandatory requirements across the EU but provides some ability for Member States to legislate locally on certain discrete matters, including the use of health data.

The GDPR applies across four main areas:



In line with the GDPR, data protection must increasingly be built into data initiatives in order to avoid severe penalties.

- This includes, for example, appointing data protection officers, conducting data protection and impact assessments, and more systematic breach reporting.
- These steps to anonymise and protect patient data therefore require time and funding to employ the right personnel and follow robust processes.
- As a result, fines for data breaches or failure to comply with the law are becoming more commonplace: the GDPR allows for data processors and controllers to be fined up to €20 million or 4% of total annual worldwide turnover for GDPR breaches.
- Overall, stakeholders feel that the balance between bureaucracy and deriving insights from data is not adequate – this is unlikely to improve given the new GDPR requirements, but GDPR will raise the bar and the quality of data sources that are able to comply.

Several actions can be taken to ensure GDPR compliance and to limit potentially negative impacts of the new regulation on oncology health data, but bear into account that this does not constitute legal advice. Legal counsel should be sought to ensure GDPR compliance.

# Recommendations

**Data collection**

Particular care must be taken when purchasing analysis and acquisition devices. As data producers, they must allow file generation in open (i.e. non-proprietary) and documented formats. Moreover, it would be easier to have an alignment of the technical possibilities of the different materials on the SUDOE space. This would allow acquisitions with the same characteristics to be made, thus avoiding the need to pre-process data at the very beginning of the chain when using data from different sources.

**Data storage**

First and foremost, the physical security of the data must be ensured. This requires the use of secure and redundant data centers. In France, security is validated by the HDS approval, which certifies providers able to provide a service for health data.

Work must be carried out to harmonize the use of data by proposing common ontologies to collect data from different regions of the SUDOE space.

In the same way, important work must be done on the definition of common metadata that will allow a good description of the bases for making information available.

Finally, to enable long-term use of the data, it will be important to ensure their durability by setting up an archiving system that takes into account both storage aspects and the software required for their use.

**Data models**

In other to query and to analyse different datasets, it is necessary to use an interoperability model based on a Common Data Model (CDM). More precisely, we propose to use the Observational Medical Outcomes Partnership (OMOP) as Common Data Model (CDM). This CDM allows to analyse of disparate source databases. OMOP permits to transform data contained within source databases into a common data model based on a common representation (terminologies, vocabularies, coding schemes). Using this common model has the advantage of performing analyses using a library of standard analytic routines that have been written. The OMOP CDM is promoted by the Observational Health Data Sciences and Informatics (OHDSI) which is an international community of stakeholders committed to bringing out the value of health data through large-scale analytics. The last version of OMOP CDM (5.0.1) is detailed in https://www.ohdsi.org/data-standardization/the-common-data-model/.

**Data processing chains**

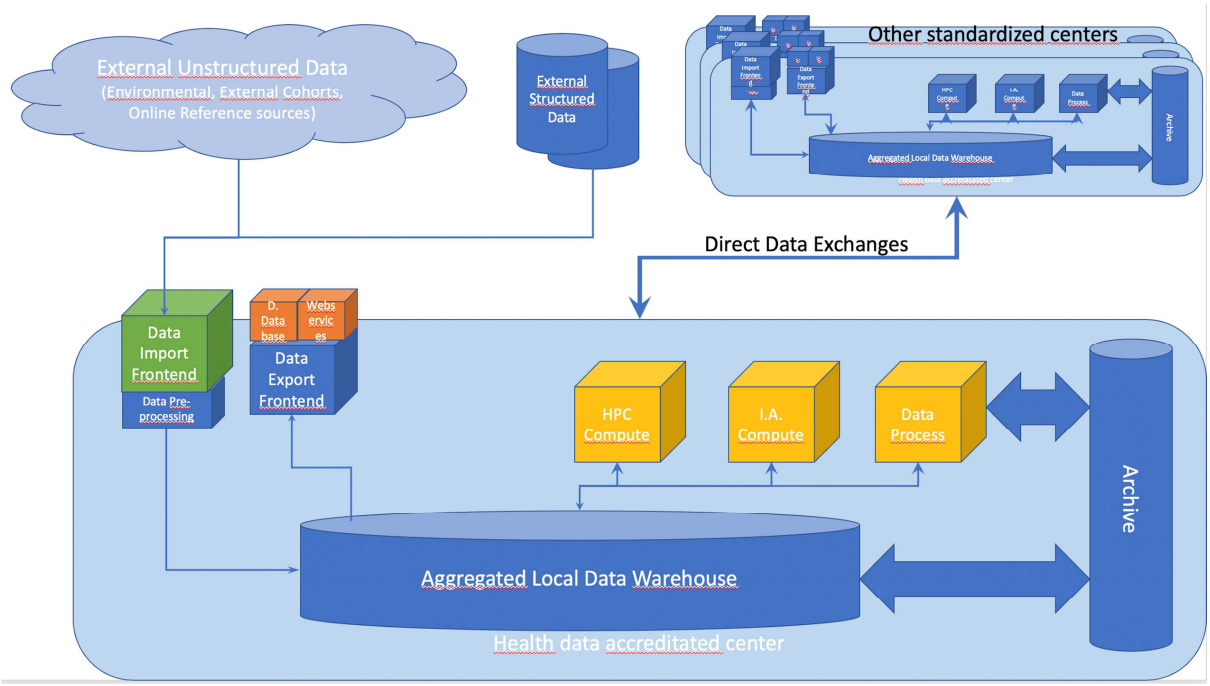
Data processing chains must be created using a modular architecture that allows for independent component implementation. This architecture will permit, through the use of intermediate formats, both the aggregation of data from different sources and the unified reuse of pre-processing in different centres.

**Dissemination and sharing**

In order to allow a good exploitation of the data present in the different environments, documented web service architecture must be proposed.

This architecture must allow both a continuous supply of data warehouses, a consultation with secure access rules and an easy link with processing infrastructures. A support team specialized in the processing of medical data must also be proposed in order to simplify the use of the various resources.
With regard to processing infrastructures, it should be noted that it may be interesting to have different ECU architectures around the same data warehouse. Indeed, with the emergence of deep neural network treatments, it may be useful to have GPU-type processing capacity in addition to more traditional HPC-type resources.

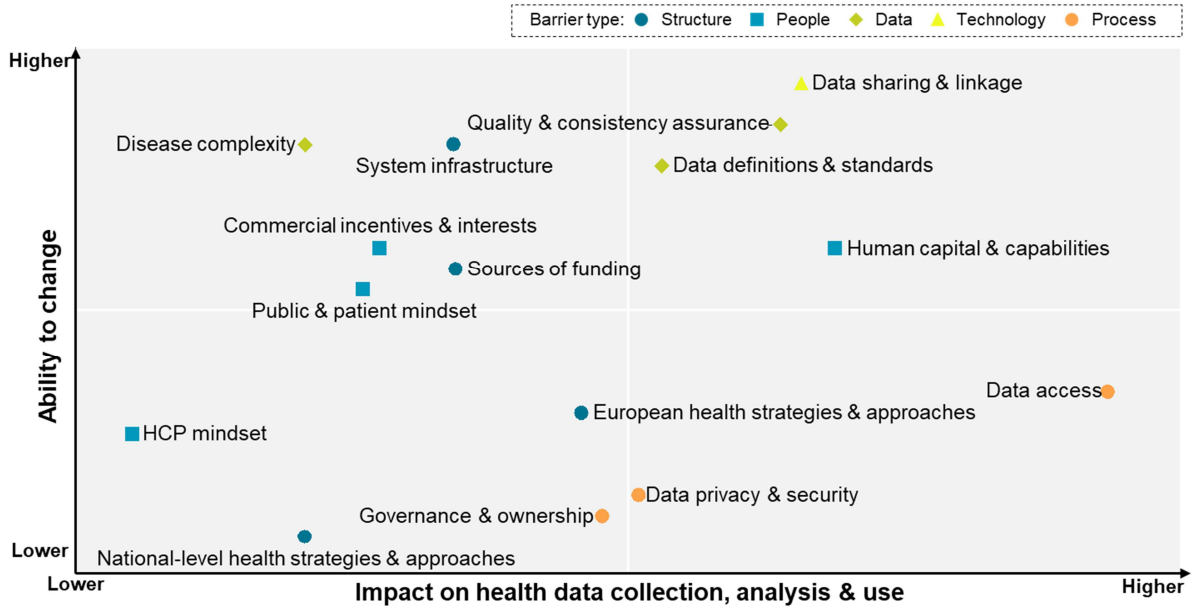
**Typical recommended architecture:**




The above diagram illustrates an ideal vision of what could be implemented in the various oncology data processing and management centres.
Data management, based on upstream standardization of formats according to the data models standard set out above, will be carried out in a unified and standardized manner between all centres. This standardization will allow a direct exchange between these data centers for, for example, making comparisons or composing virtual cohorts. In each of the centres, different means of information processing can be found that will allow flexibility on processing possibilities and elasticity of computing capacities. A permanent archiving system that will manage cold data (i.e. not subject to processing) will be set up in each centre (or possibly shared). This archiving system must allow both data storage and monitoring of the software necessary for their use.
In order to exchange with ecosystems outside this network of centres, pre-treatment and export machines will be available at the entrance and exit. Pre-processing machines will ensure the standardization of input data and the availability of data from other environments (weather data, consumption, etc.). The export machines will allow data to be exposed both in proprietary environments and in an open science perspective. These export machines will also offer a service to document the data present in the centres and the possibilities of use.

Conclusions

Although many of these challenges are significant, several (e.g. technologies, mindsets) have seen positive changes in recent years and have been overcome in various countries. Others remain that may require limited effort to improve rapidly, for example around establishing quality assurance or the right data infrastructures. Still others, such as data access, data privacy and security, and European health strategies, are anchored in processes or legislation and may take years to evolve. On the whole, however, there is greatest potential in tackling the challenges that are strongest but also more open to change, such as data definitions, standards, sharing and linkage, and building capabilities – this will require long-term, forward-looking collaboration across all stakeholders who stand to benefit from health data.




Improvements in overall data granularity, quality and interconnectivity are particularly necessary in oncology due to the increasingly stratified nature of the disease. Guidelines and new technologies are paving the way to improving this, but further alignment and understanding of the use of health data to improve care, incentives to build this data and scaling ability are still required.

Moving forward, a new approach is required that focuses on providing sustained funding and enabling the linkage of datasets to provide a 360 view of disease, treatment dynamics and the patient experience, while respecting the need for confidentiality. Local legislation and infrastructure may limit the ability to do this, but examples of where this has been successful can challenge established mindsets and open up systems to the use of RWD.

Processes are already complex and burdensome, and likely to become so given the understandable GDPR push for transparency and better use of health data. Better planning and systematic efforts to put patients at the center of data collection and use in a user-friendly way, most likely using new technologies, represent the best options to simplify this environment.

GDPR will raise the bar in terms of the quality and management of health data, with requirements likely driving professionalisation of data collection activities. Larger datasets will emerge from this, bolstered by new technologies that must be embraced, and their use accelerated.

The lack of data science skills and ability to retain talent in healthcare currently limits the quality of RWD available, which is particularly challenging for oncology given the complexity of the disease. Mentalities are changing, however, and with better information and communication around health data, cancer communities can play a greater role in sharing data and ensuring patients benefit.

# References

Gesulga Jaillah Mae et al., Barriers to Electronic Health Record System Implementation and Information Systems Resources: A Structured Review, 2017
European Federation of Pharmaceutical Industries and Associations (EFPIA) Report on oncology health data in Europe
Hogan Lovells, The Final GDPR Text and What It Will Mean for Health Data, 2016;
Taking advantage of patient data – an outlook on the upcoming General Data Protection Regulation, 2017
https://www.hldataprotection.com/2016/01/articles/health-privacy-hipaa/the-final-gdpr-text-and-what-it-will-mean-for-health-data/
https://www.noerr.com/en/newsroom/News/taking-advantage-of-patient-data-an-outlook-on-the-upcoming-general-data-protection-regulation.aspx