scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A comprehensive European Colorectal Cancer Cohort dataset

Petr Holub *et al.*#

Colorectal cancer (CRC) is a leading cause of cancer-related deaths worldwide. The Biobanking and Biomolecular Resources European Research Infrastructure (BBMRI-ERIC) established a CRC-Cohort with European coverage contributed by 26 biobanks from 12 countries. This retrospective, multi-center study contains structured and curated clinical data, supporting research on biomarkers for early detection, prognosis, and treatment. A phenotypical/clinical data model has been defined and individual-level data from 10,780 CRC patients have been collected at BBMRI-ERIC in the central data deposition service hosted as part of its services. The participating biobanks host additional data, which can be accessed on request and used to derive additional data. This mechanism has been used to extend the collected data with scans of histopathological slides to support research in artificial intelligence in digital pathology and with whole genome sequencing data to pilot a use case of the upcoming European Health Data Space (EHDS). Here we present the methodology, the quality assurance mechanisms, and the implementation of FAIR and FAIR-Health principles applied to build the CRC-Cohort.

Background & Summary

According to Global Cancer Statistics 2020[1], Colorectal cancer (CRC) is the third-deadliest and most commonly diagnosed cancer worldwide. Its global incidence and mortality have increased over the last few years[1] and are projected to reach 3.2 million cases annually by 2040[2]. This makes CRC a major global health challenge requiring innovative solutions for risk prediction, early diagnosis and targeted therapies[3]. Hence, early diagnosis of the pre-malignant lesions from which CRC develops, is key to combating the disease. The slow progression from these lesions, provides an opportunity for population surveillance and screening, enabling detection of CRC at early stages[4].

Currently, various screening strategies are available for CRC. The most widespread is the faecal immunochemical test (FIT), due to its non-invasive nature, simplicity and affordability[5]. More sophisticated procedures involve DNA analysis of samples from different sources[6,7], with CRC biomarkers playing a crucial role in enhancing accuracy and sensitivity. However, despite extensive research and numerous ongoing studies exploring new biomarkers, only a few have been incorporated into clinical practice[8]. One example is Kirsten Rat Sarcoma viral oncogene (KRAS), a gene that is mutated in 35% to 45% of CRC cases[9]. The presence of KRAS mutations can indicate an increased cancer risk and can guide treatment decisions. Another example is Hereditary Non-Polyposis Colorectal Cancer (HNPCC), or Lynch syndrome, a genetic disorder that increases the risk of colorectal cancer and other cancer types[10]. It results from defects in DNA mismatch repair genes, predisposing individuals to early-onset CRC.

There are also screening strategies that do not depend on biomolecular analysis or detection of gene mutations. Imaging procedures such as endoscopy[2], while usually invasive, are the most accurate tools for diagnosis of the disease[4]. Other strategies leverage histopathological samples, such as histological sections or whole-slide images (WSIs). Pathologists traditionally evaluate these samples to provide detailed tumor descriptions; the growing volume of digitized biopsies, coupled with technological advancements, has led to the integration of artificial intelligence (AI) to support their work[11,12] and is paving the way to find new biomarkers[13].

Generally, biomarkers are used alongside clinical examinations, imaging and pathology for early disease detection, predicting outcomes, and selecting targeted therapies in the era of precision medicine. For instance, stage II patients stratification exemplifies a critical medical need of healthcare optimisation, as evidence remains insufficient to determine whether surgery alone is adequate or if chemotherapy is necessary to prevent

#A full list of authors and their affiliations appears at the end of the paper.

recurrence. From a biomolecular perspective, biomarkers are footprints of the mechanisms active during CRC carcinogenesis, as the case of microsatellite instability (MSI). Other mechanisms involved in CRC progression are chromosomal instability (CIN) and CpG island methylator phenotype (CIMP)[14]. However, new pathways are continuously analysed and identified as players in the malignant progression[15]. Hence, to better understand the disease, other research targets are needed beyond biomarker discovery for clinical applications to provide a deeper look at the molecular causes of the disease – which could also help to stratify patients and bring new therapy opportunities. One of the main limitations to pursuing such targets is the scarce availability of biosamples or the lack of large cohorts of validated and well-structured clinical data, needed to avoid potential biases and have enough statistical power. The Colorectal Cancer Cohort Dataset (CRC-Cohort Dataset) dataset that we present here contributes to overcome this need, integrating clinical data of 10,780 patients with C18.0-C18.7, C19 and C20 diagnoses (in ICD-10 terminology, https://www.who.int/classifications/icd/icdonlineversions/en/) from 26 biobanks in 12 European countries.

Overall, this cohort collected by the Biobanking and Biomolecular Resources European Research Infrastructure (BBMRI-ERIC) and its federated network of partner biobanks should foster research in the CRC field, facilitating, among other applications, the discovery of new biomarkers and optimization of patients treatment. The centrally collected data is hosted on BBMRI-ERIC's data deposition capacities and, due to sensitive personal health data, it is accessible via BBMRI-ERIC expedited access procedure for centralised datasets described in this paper. These centralized deposition services are endorsed by BBMRI-ERIC Member states as a part of the governance procedures. Additional data from the originating biobanks is accessible via the standard BBMRI-ERIC access procedure.

Methods

Data collection. The collection of the CRC-Cohort Dataset data was part of the ADOPT BBMRI-ERIC project (https://cordis.europa.eu/project/id/676550) by BBMRI-ERIC and partner biobanks. In this process, several challenges had to be addressed, such as how to overcome different data quality and data structures, finding a common legal basis to make this data available in compliance with the General Data Protection Regulation (GDPR), its heterogeneous implementations, as well as varying ethical requirements. The planning of the work started in 2016 and comprises a number of activities briefly summarised here.

The CRC-Cohort Dataset data was collected in the central hub by a BBMRI-ERIC judicial person. The data model was developed through a consensus-driven process involving clinicians, IT experts, and biobank representatives. This approach emphasised unambiguous definitions to ensure interoperability across systems. Following repeated iterations of review and refinement, the final version of the data model was finalised in 2018 and implemented into the software that enabled biobanks to input their data in a standardised format.

For inclusion in the cohort, specific criteria were established, focusing on cases with primary CRC, surgical materials (excluding biopsies), and the availability of mandatory data. To ensure the utility of the dataset for medical research, cases with at least five years of follow-up data were prioritised, providing clinically relevant survival data.

A comprehensive Data Protection Policy (DPP) was developed to align with the GDPR. This policy addresses data governance, access mechanisms, and privacy protections. BBMRI-ERIC assumed the role of data controller for the cohort, except in Finland, where it acted as a processor due to specific legal constraints under Finnish law. This distinction allowed Finnish biobanks to contribute while complying with national legislation.

The data collection process began with recruitment efforts targeting biobanks across Europe. A significant technological infrastructure was implemented to support the data collection. A centralised Metadata Repository (MDR) was established to manage the data model in a machine-readable format, ensuring consistency and FAIR compliance. Tools for data harmonisation and quality checks were also developed to support the conversion of various data formats into the structured XML required by the centralised system. In addition, the Colorectal Cancer Data Collection (CCDC) system was created, which enabled centralised data collection, including database, web-based user interface for data entry and management, and a secure REST API to allow programmatic imports of the data. The central CCDC system for collecting data was implemented based on open-source software from the OSSE Project[16]. The web-based user interface was designed for the small biobanks contributing data manually only. It was also possible to add or correct missing data via the web user interface for the biobanks contributing the initial data via upload, but this method was discouraged for reasons of very complicated maintenance of this hybrid upload/manual data entry. The REST API uses XML data payload. Two XSD-based data validation schemes have been developed: one for full strict validation, which detects also any missing data required by the data mode, and one for partial validation, which does not take completeness into account and focuses on validating the provided data. The latter mode allows biobanks to upload incomplete data, obtain feedback from the data quality checks, and subsequently improve the dataset handed over to BBMRI-ERIC. For easier understanding of the format, the biobanks were also provided an example of synthetic data in XML format. The whole package is available via Zenodo[17].

The system was deployed and hosted on the BBMRI-ERIC production IT infrastructure managed within BBMRI-ERIC Common Service IT. For contributing biobankers it was available using the secure HTTPS protocol over public Internet. For administrators, who have access to management operations, it was only accessible following successful two-factor authentication (first factor being VPN authentication required to reach the service's administrative interface; the second being authentication using local server accounts via the SSH protocol).

The design and implementation of data harmonization tools was needed to support the conversion process from common tabular data (Excel, CSV, TSV, etc.). The CRC-Cohort Dataset opted for complex data formats, such as deeply structured XML, to fully represent the dataset including the semantic relationships between entities. These include, for example, relationships between treatments and responses to treatments. However,

most of the data used by the biobanks were only available in simple tabular formats, such as Excel, CSV, or TSV. Similarly, many biobanks lacked advanced IT systems and the necessary skills to transform their data into complex target formats, such as the XML format required by the API of the CCDC system. Transforming such data is challenging, particularly because in addition to ensuring the correct syntactic transformation of the data, care must taken to properly map equivalent semantic entities from the source to the target data model. To support this process, a dedicated toolset[18] was developed which accomplishes this task via a semi-automatic mapping and a fully automatic conversion of the data into the XML target format, including initial data quality checks. The system also maintains a database of mappings that can be reused when either new data arrive from the same biobank or when starting with a new biobank from a similar national/regional context. Finally, the toolbox was largely operated directly by the central IT team, as the biobanks were essentially interested in providing data in their own formats. However, there was a lively exchange of information with the biobanks to ensure correct interpretation and assignment of values.

The project faced several challenges, including disparities in IT capabilities and legal constraints among the biobanks. The design and implementation of data quality checks in collaboration between expert pathologists and IT experts, was a key point, as described more in-depth in the Technical Validation section. Iterative cycles of data quality improvement were conducted in collaboration with domain experts to address these issues. While some biobanks expressed concerns about the burden of repeated revisions, compromises were made to ensure high-quality data without excessive demands. The implementation of new data quality checks and the consequential submission of improved data was repeated until either all problems were fixed, or the biobanks acknowledged that the detected problems were false positives, or they were no longer able to improve the quality of the data. Some biobanks complained that the iterative nature of this process posed a substantial burden for them and wanted to contractually cap the number of data quality check iterations. While this may initially appear to be a reasonable request, it actually creates two problems: (a) a new upload of the data after fixing some problems can trigger another positive problem report—e.g., by a new check, by a check that was previously unable to run (despite that during the implementation we tried to report as many problems at once as possible—sometimes at the cost of reporting two or more different interdependent problems as separate reports that were fixed by correcting single data issues only); and (b) new data quality checks were being developed even after the initial implementation phase. A compromise was typically found with a focus on fixing the most problematic data in the first rounds.

As a result of the data collection process, the CRC-Cohort Dataset has been registered in BBMRI-ERIC Directory (bbmri-eric:ID:EU_BBMRI-ERIC:collection:CRC-Cohort, http://hdl.handle.net/21.12110/1.adc0-d6ce-bcfb) and in FAIRsharing.org (https://doi.org/10.25504/FAIRsharing.N4a3Pj) and made queryable in a privacy-preserving manner using the BBMRI-ERIC Federated Platform.

From that point in time the work on the cohort did not stop, as continuous maintenance and improvements are ongoing. One of this efforts is the establishment of a cBioportal instance and the integration of a WSI visualisation tool.

Inclusion criteria of CRC-Cohort Dataset. The following consensus was reached on the inclusion criteria (not directly part of the data model, but also necessary for correct interpretation of the resulting dataset):

- Colorectal cancer as a primary diagnosis (C18.0 to C18.7, C19, C20).
- Available FFPE—i.e., surgical material.
- Availability of all mandatory data as defined by the model.
- Willingness to provide access to (a) samples, (b) pseudonymized data as a part of (i) participation in research projects, (ii) cost or no-cost recovery procedure (assuming prior signature of material/data transfer agreement).

Note that biopsies do not qualify as surgical tumor material, as they do not provide sufficient material to support multiple research projects.

It was *recommended* to focus primarily on cases with at least a five-year follow-up for obtaining clinically relevant survival data[19–21]. As the CRC-Cohort was developed until March 2019, the available cases were mostly diagnosed and the samples collected until beginning of 2016 (note that of dead patients, shorter than 5 year period was possible).

Cohort expansion. Recently, the CRC-Cohort Dataset was expanded through the addition of associated whole genome sequencing (WGS) data of 425 patients from the Uppsala biobank[22], enriching the baseline information about the status of the most important mutations for the CRC available for all cases. This expansion was boosted by BMRI-ERIC's participation in the HealthData@EU pilot project (https://ehds2pilot.eu/), which aims to pilot the HealthData@EU infrastructure for the secondary use of data envisioned in the European Health Data Space (EHDS) regulation proposal (https://www.consilium.europa.eu/media/70909/st07553-en24.pdf) Similarly, the involvement of BBMRI-ERIC in another project, EOSC4Cancer (https://eosc4cancer.eu/), led to the employment and customization of a cBioPortal[23–25] instance for the cohort, where statistical analyses of aggregated clinical and genomics data can be performed and presented together with the corresponding WSIs. This functionality facilitates data access while maintaining security. A link to xOpat[26] has also been implemented, enabling the use of this tool for whole-slide image visualisation. Security measures, such as VPNs, HTTPS protocols, OAuth2, and two-factor authentication, are implemented to ensure secure data access and handling. Access is managed through a structured authorisation and authentication process.

Data model development. Based on nominations from BBMRI-ERIC National/Organizational Nodes, a core interdisciplinary working group including biobankers, clinicians, disease registry experts, researchers, and IT experts was organized.

This working group designed the model of the data to be collected by repeatedly iterating the following process: (*a*) acquisition of basic consensus on inclusion criteria; (*b*) acquisition of basic consensus on collected attributes among the medical experts; (*c*) development by IT experts of formal model including entities, their attributes and their mutual relations; (*d*) review of the formal model by entire working group; and (*e*) approval of the resulting formal model by the BBMRI-ERIC Management Committee (used also as the project management board in ADOPT BBMRI-ERIC project)

The activities were organized through a series of teleconferences involving the entire working group, supported by breakout meetings involving specialized subgroups to deal with specific issues (e.g., oncologists describing possible treatments and outcomes of treatments, molecular biologists defining the most relevant genetic variations). The process was completely consensus driven: there were no items that could not be resolved during this design phase.

The initial model was presented to the biobanks, which had the opportunity to review it. This resulted in a review document that was maintained and the data model was further refined. The entire process was concluded in January 2018: at that point, the model was declared immutable and its final version was implemented into the software for collecting the cohort. Based on the final model, instructions for biobanks were prepared, detailing how to format input data for import into the central database.

The resulting data model, described in more detail in Section Data model implementation, prioritizes the unambiguous definition of the data structure so that it can be implemented in IT systems, and defines which parts of the data model are required to obtain data sets meaningful for medical research.

Conversion of the clinical data to the openEHR format. Following the original collection process, clinical data has been converted to openEHR[27], one of the main formalisms used in healthcare to capture clinical information while including its semantics. The openEHR framework enables open and structured data representations, decoupling the meaning of the data from how they are persisted, thanks to the adoption of abstract models (the clinical archetypes) defined in the community-maintained openEHR Clinical Knowledge Manager (CKM). Archetypes represent the extended description of a single clinical concept, including all possible attributes, and are combined in templates, to express the specificity of a use case, including limitations to possible values assumed by the data elements or bindings to ontologies and terminologies. Thus, this representation of the CRC-Cohort Dataset helps to precisely preserve its semantics in a manner understood by the healthcare community, thus promoting both its interoperability and reusability.

Therefore, a mapping was created from the CRC-Cohort Dataset data model to a new openEHR model (CRC_cohort template, https://github.com/crs4/crc_cohort_modelling) designed specifically to accommodate this data. After a preliminary selection of possibly relevant archetypes from the openEHR Clinical Knowledge Manager, the different attributes of each element of the source model—i.e., meaning, datatype, cardinality, allowed values, etc.—were analysed to find a proper correspondence with archetypes' nodes. Archetypes were then assembled to form the template. When defined, mapping descriptors were embedded as "annotations" in the template, for each node and for each assumable value, in the form of key-value pairs that could be processed automatically during conversion.

To support semantic interoperability, we leverage the possibility for openEHR to bind external terminologies to data nodes and values. Therefore, model and data instances are further enriched by associating concepts from the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) (https://ohdsi.github.io/CommonDataModel/, available via the ATHENA repository: https://athena.ohdsi.org/). These mappings of the possible values exist both in the template and the final data instance. In fact, at the template level, it is possible to restrict the possible values for a node to a predefined list, where OMOP concept IDs can be added and used afterwards in the data instance. The detailed mapping of each element of the CRC-Cohort Dataset data model and XSD to the openEHR template and to OMOP is available on GitHub (https://github.com/crs4/crc_cohort_modelling/blob/main/documentation/openEHR_mapping_documentation/CRC-Cohort_openEHR_mapping_documention_v1_Jan_2023.pdf).

The structured and semantically well-defined openEHR dataset is an ideal source for converting the CRC-Cohort Dataset into other formats. Starting from this source, we have converted the dataset to OMOP, enabling searches through BBMRI-ERIC Finder (https://finder.bbmri-eric.eu/) – which is part of BBMRI-ERIC Federated Platform (https://www.bbmri-eric.eu/bbmri-sample-and-data-portal/). This integration allows privacy-preserving querying of individual-level and sample-level data through the Federated Platform producing obfuscated counts of matching patients. The same openEHR source was also used to map and convert the dataset to HL7 FHIR[28] (https://hl7.org/fhir/). A software tool was developed (https://github.com/samply/bbmri-fhir-gen) to generate FHIR test data following the BBMRI-ERIC Implementation Guide (https://samply.github.io/bbmri-fhir-ig/, source code available at: https://github.com/samply/bbmri-fhir-ig). By converting the data to FHIR, it is also made searchable through the BBMRI-ERIC Locator (https://locator.bbmri-eric.eu/) – another component of the Federated Platform. However, to avoid any confusion from overlapping query functionality provided by the two systems, the dataset is disabled in the Locator and is only visible in Finder results, where more data can be queried.

Conversion of the clinical data to OMOP. The openEHR fields are expressed using OMOP CDM. This mapping relates the openEHR data model for the CRC-CohortDataset, expressed in the form of a template, to the elements of the OMOP Common Data Model, v5.48. The mapping has the OMOP CDM tables as a starting point. For each table element, an analogous element was sought in the openEHR model that was similar in terms

of data type, valid values and semantics. We define the mapping of each element as "valid" when these aspects are considered to be sufficiently overlapping or when it is possible to adapt the source element (openEHR) to the intended form of the target element from the CDM table. However, given the difference in the content structure of the two approaches, there are many special cases for which a direct mapping is not so obvious or even not possible (e.g., versions of the World Health Organization (WHO) and Union for International Cancer Control (UICC) standards used to determine tumor grade and disease stage, or elements from molecular markers, such as instability of specific microsatellites, or HNPCC risk situation based on Amsterdam criteria). Of note, many OMOP CDM tables, except for "Person", have elements to express the provenance of the reported data (e.g., period_type_concept_id! in the Observation Period table). Considering that the data has been collected within a research project, we chose the Case Report Form option (Concept ID: 32809) among all the OMOP accepted concepts. This assumption is maintained in every table that requires this information.

Given the differences between these two models, mapping is an on-going process, which is recurrently being validated by domain experts to support both the sharing of the mapping/conversion approach and the validation of the mapping/conversion itself.

Accessing data through cBioPortal. cBioportal[23–25] is a web application for statistical analyses of clinical and genomic data. It allows third-party resource integration using HTML iframes, making possible to integrate the xOpat viewer (https://github.com/RationAI/xopat, also available at: https://xopat.readthedocs.io/) for showing WSI slides related to those data. Users can request access to individual studies, that—after approval—appear in the study overview. In cBioportal, they can view available data presented in graphs, and optionally follow a resource link to the WSI viewer xOpat, which presents the data read through WSI-Service (https://github.com/RationAI/WSI-Service/). Regarding Authentication, Authorization & Security; third-party resources in cBioPortal are simple: a study can specify a resource URL the cBioPortal uses to display additional data about a sample. However, user authorization is much more challenging. CBioPortal does only support study-level access granularity (cBioPortal also supports study *views*, but these are user-centered; users can share them with other people via URLs.). Furthermore, it is necessary to control all services through LifeScience Login (also known as LifeScience Authentication and Authorization Infrastructure, AAI, https://lifescience-ri.eu/ls-login/), OpenID Connect (OIDC) provider.

- *Secure Deployment*. The data is hosted on a Virtual Private Network (VPN)-protected *data server*. The *application server* hosts cBioPortal, xOpat, and acts as a reverse proxy to WSI image server, hosted at the *data server*. The *application server* uses HTTPS protocol, and has a VPN tunnel to the *data server*.
- *User Authentication*. Users can log-in through LifeScience Login using their university, hospital and similar identity providers. Access requires Multi-factor authentication (MFA).
- *Study Access Authorization*. Users are added to groups that represent individual studies. This information is propagated to both cBioPortal and xOpat. cBioPortal supports custom OIDC integration from version 6. However, user identities and their *authorities* (cBioPortal entities that encode study access) have to be stored in the internal database with custom OIDC provider integration, and does not support 2FA verification. To overcome this, cBioPortal was customized to allow for third-party OIDC authorization configuration; and this customization was requested for inclusion in the official repository. xOpat handles authorization on the WSI image server level. Both these services solve authorization by dependency injection: an interface is implemented that parses available user information provided by LifeScience Login. Of note, this approach requires management of customized builds.

Data Records

Dataset. The CRC-Cohort Dataset[29] (available at https://directory.bbmri-eric.eu/ERIC/directory/#/collection/bbmri-eric:ID:EU_BBMRI-ERIC:collection:CRC-Cohort?search=crc) includes clinical data containing:

- detailed phenotypic and clinical data related to the patients: biological sex, age at primary diagnosis, relevant genetic and other risk factors, vital status and survival information, treatment information and outcomes of treatment;
- metadata about biological samples with description of morphological features;
- associated digital pathology reports and histopathological diagnosis, including: pathology-derived TNM ("Tumour", "Nodes", "Metastases") Classification of Malignant Tumours (pTNM), UICC staging[30], WHO grading, and ICD-10 classification and status of most relevant molecular markers (KRAS and NRAS for exons 2/3/4, microsatellite instability, status of HNPCC risk based on Amsterdam criteria, and mismatch repair gene expression).

Moreover, the dataset includes WSIs and genomic data. The centralised sub-cohort contains 3,260 WSIs of formalin-fixed paraffin-embedded (FFPE) slides of colon tissue samples for 1,433 patients scanned at 20 × or 40 × magnification, totaling 26 TB of primary data contributed by 9 biobanks from 8 different countries. The WSI data is stored and available in the original file formats: Mirax format (MRXS), Aperio format (SVS), and Hamamatsu format (NDPI), all readable via OpenSlide library[31]. In addition the images have been converted to OME-TIFF,[32–34], an open, accessible, multi-dimensional, multi-resolution format that enables fast access for computational processing and provides flexible metadata storage. Genomic (WGS) data is available in Variant Call Format (VCF) files containing somatic mutations of 425 patients from the Uppsala biobank.

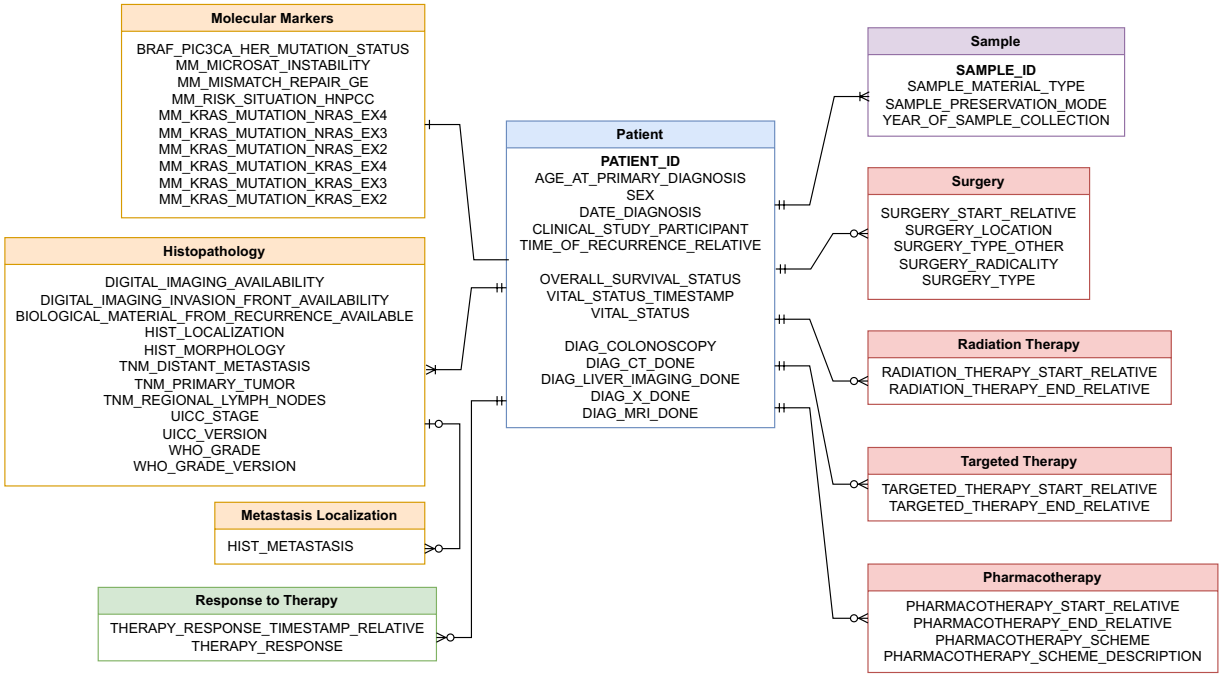

Fig. 1 Simplified entity-relation diagram of the CRC-Cohort Dataset data model. Note that this does not reflect details of the actual database implementation such as using primary keys and normalization of the database.

Data model implementation. The original primary storage of clinical data is in a data management system based on OSSE[16,35], which uses a relational PostgreSQL database with hierarchical JSON data structures directly corresponding to the data model defined in the MDR,[35,36], compatible with ISO/IEC 11179, in machine-readable form to be used by applications. It is a common approach in the medical informatics domain that the data models are implemented in a dedicated MDR. This approach allows the data models to be more easily reused in multiple systems and helps avoid having different systems become incoherent in light of future model updates. However, the practical implementation of the model revealed a common design problem: the data model is often documented on the level of attributes of entities, while the relations of the entities are part of the design of the target system[36]. This includes the fact that (at least simple) attribute constraints can be part of the MDR while relations on cardinalities of relations can not. Hence the data model coming from MDR is incomplete and needs to be complemented with additional information in order to enable biobanks to contribute practically. The MDR instance hosting the data model is publicly available (https://mdr.osse-register.de/view.xhtml?namespace=ccdg). Availability of the data model in a machine-readable structure is a prerequisite for the system to be FAIR compliant. The data model is implemented in XML/XSD[17] and the specific number of files or tables accessed by the user when the data access is granted depends on each request, according to data minimisation principles. Similarly, the access format depends on the user's request, as clinical data can be provided as plain text in different formats (e.g., openEHR Canonical, FLAT or Structured JSON serializations: https://specifications.openehr.org/releases/ITS-REST/Release-1.0.2/overview.html#header-json-format; openEHR Canonical XML serialization: https://specifications.openehr.org/releases/ITS-REST/Release-1.0.2/overview.html#header-xml-format; OMOP CDM tables in CSV format: https://github.com/OHDSI/CommonDataModel/tree/main/inst/csv).

Data was pseudonymized before collection—i.e., the data is linkable back to the original individual cases at the source BBMRI-ERIC partner biobanks, where direct identifiers have been replaced by non-speaking non-reusable codes. Requirements on pseudonymization are defined in the DPP and implementation is responsibility of the source biobanks.

The data model of the centralized data collection has been designed as a medium-depth clinical model, with the primary aim to support the discovery of cases (patients) relevant for clinical research scoped as outlined in the Background and Summary section above. The model comprises several entities, which are schematically illustrated in Fig. 1. The central entity is the Patient, which is linked to the Histopathology, Molecular Markers, and Sample entities. Several types of treatment relevant for CRC are defined – including Surgery, Radiation therapy, Targeted therapy, and Pharmacotherapy – while the progress of the disease is represented using the Response to Therapy entity. Cardinalities of these relations are shown in the diagram.

The model also encompasses information on WSIs and genomics. WSI metadata include information on the case (patient) and staining. There can be multiple WSIs available per patient, for different histological slides and possibly different histological stainings. Genomic data is included in the model by linking VCF files containing somatic mutations to the other data elements in the cohort.

For each of the entities, the tabular view of data model (from page 39 of https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bc49b2c0&appId=PPGMS) gives a detailed description of each of the collected attributes. Note that in some cases, such as Molecular Markers entity, there

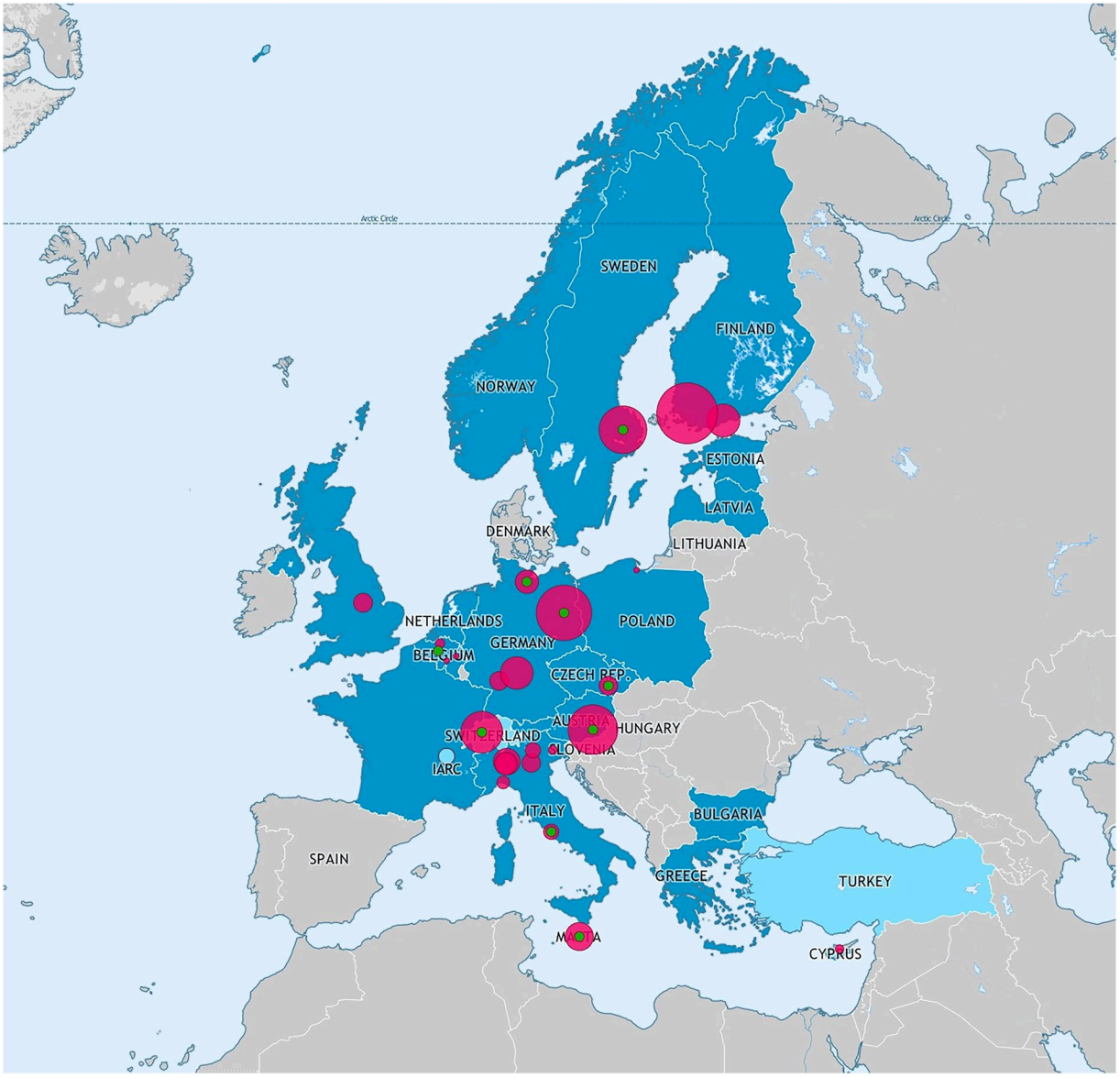

Fig. 2 Map of BBMRI-ERIC partner biobanks contributing to CRC-Cohort Dataset. Each contributing biobank is marked as magenta circle with size of the circle proportional to the number of contributed cases. A green dot in the middle of the circle marks biobanks that also contributed WSIs. The coloring of countries reflects state of membership in BBMRI-ERIC at the time that the CRC-Cohort Dataset was collected: dark blue are members, light blue are observers.

are attributes that are required, yet there is a "not done" (or similar) option among the list of permitted values. The reason is that if the given attribute was optional, it could have been omitted for different reasons, while in this structure it has to be explicitly stated that given data is not available.

Finally, and as explained in the Methods section, following the original collection process the clinical data has been converted to openEHR[27]. From there, clinical data is mapped and converted to the HL7 FHIR standard and to OMOP CDM.

Data Overview

The CRC-Cohort Dataset comprises approximately 70 TB of data, containing a total of 10,780 cases and patients (one case per patient). It integrates clinical data of patients with C18.0–C18.7, C19 and C20 diagnoses (in ICD-10 terminology), covering UICC stages I, II, III and IV, with a focus on stages II and III. This centralised dataset contains data from 26 biobanks in 12 European countries, providing good geographical coverage of the entire European continent. Geographical distribution of contributing biobanks is shown in Fig. 2 and per-biobank contribution statistics are presented in Tab. 1. The cases are mostly from the year 2014 or earlier, due to the 5-year follow-up recommendation, but some biobanks also contributed cases with shorter follow-up information. Overall, the CRC-Cohort Dataset is based on a robust framework for harmonising biobank data for medical

Biobank	Cases	WSIs
Auria Biobank	1344	
Central Biomaterial Bank Charite	1226	21
Biobank Graz	1066	2214
Uppsala Biobank	1017	
Gewebe Biobank - Bern	871	532
Helsinki Biobank	653	
ibdw Wuerzburg	646	
Malta Biobank	533	
INT Biobank	500	
Interdisziplinäres Centrum für Biobanking-Lübeck	417	205
Humanitas Cancer Center	397	
Nottingham Health Science Biobank	308	
Masaryk Memorial Cancer Institute	300	268
Deutsches Krebsforschungszentrum	300	
ARC-NET Biobank	300	
Cancer Center Regina Elena Biobank	218	20
TrentinoBioBank	218	
Centro Risorse Biologiche CRB-USMI	155	
Biobank Antwerpen	69	
CRO BIOBANK	55	
CING Biobank	50	
Biobanque du laboratoire d'anatomie pathologique	50	0
Medical University of Gdansk	47	
CHU UCL Namur	20	
CHU Brugmann	10	
Biotheque Hospitalo Universitaire de Liege	10	

Table 1. Statistics of per-biobank contributions to the CRC-Cohort Dataset.

research. Despite significant organisational and technical challenges, it provides streamlined data-sharing mechanisms while adhering to legal and ethical standards, setting a benchmark for future data collection initiatives.

Technical Validation

The responsibility for data quality is a contractual obligation on each source BBMRI-ERIC partner biobank. The central quality assurance framework of CRC-Cohort Dataset has been implemented in the following steps.

1. BBMRI-ERIC implemented basic automated validation on the data that was entered or imported into the CCDC system. The implementation uses the partial XSD-based validation, so that syntactically correct (conformant) incomplete data can be provided and gradually completed. This approach provides immediate feedback to the source biobank when providing the data.
2. BBMRI-ERIC implemented a more advanced quality analysis system, which checks data completeness and plausibility. This analysis is run periodically and results are made available to the source biobank.
3. Source biobanks receive quality feedback from the previous two steps, improve the data (as per their contractual obligation, and to the extent possible), and provide updated data. The previous two steps are then repeated.

XSD-based validation. The XSD-based validation checks for syntactic conformance of the data uploaded and—to some extent—for completeness of the data. Syntactic conformance means that the XML is properly formatted, if XML import of data is used as an entry. Further, this level of validation verifies whether all required elements are present, whether values are of proper data type and, in case of lists, that the value of the given variable is one (or more if allowed) of the permitted values for the given list type. XSD-based completeness checks do not check for availability of all necessary "forms", which are representing individual entities in the entity relation diagram in Figure 1, but only checks if the "form" itself is complete. The same validation, except for XML formatting, is also applied when the data is entered manually via the web-based interface of the CCDC application.

As previously mentioned in the text, even completeness checking within the single entity "form" was sometimes a hurdle for the biobanks and hence two XSDs have been made available: strict and relaxed; the relaxed XSD ignores any completeness checks and allows iterative improvement of data completeness over time. In addition to having integrated this validation in the CCDC data input flow, the XSDs were provided to the biobanks to allow them to perform checks locally/offline when preparing data for import.

Advanced quality analysis. The advanced quality analysis framework has been implemented to provide near real-time feedback to biobanks on completeness and plausibility of data they provided. Instead of creating hard errors which would prevent from data being saved into the system, the advanced quality checks do not prevent users from entering any data and rather guide them to improve data quality. Further, not all checks necessarily test for serious errors. Some checks produced soft warnings which only highlighted suspicious data which had some probability of containing an error and warranted being verified.

The quality checks were implemented centrally on the CCDC server using R[37]. The checks are run on the CCDC server automatically in a periodic manner and the results are sent back to the respective biobank after they contributed with new or updated data. The reports are produced as Excel files, since this format is the most approachable for the biobanks, and they are made available for secure download. This feedback is typically followed up by consultations with the biobanks to interpret the data quality check report correctly and to provide guidance towards corrective measures. The following areas are covered by these advanced quality checks.

- Date-based and age-based checks
    - Diagnosis date and last vital check date are compared, raising a warning if the latter precedes the former or if both are the same date.
    - The difference between the above dates was calculated and compared with the survival time provided. These two values should match, otherwise this is also flagged as a potential inconsistency.
    - Vital status timestamp and initial diagnosis date are compared with the current date to check if they are future dates—which is undoubtedly an error in the data.
    - Age at primary diagnosis was checked, giving a warning if the patient is diagnosed at a suspiciously young age (less than 15 years old).
    - Overall survival after first colon cancer therapy started was checked, raising warnings when:
        - patient overall survival value is over 4,000 weeks ($\approx$ 77 years);
        - patients reach 100 or more years of age when summing the overall survival time and the age at primary diagnosis—the latter being less than 95 years;
        - large differences between survival and last reported follow-up of the patient, or very similar values of survival for large groups of patients (e.g., one biobank provided the same survival for all living patients).
- Check of follow-up consistency and completeness of events and follow-up
    - The start of therapy (radiation, targeted or pharmacotherapy), given in weeks since the initial diagnosis, is checked and verified to be lower than the end of these events (also indicated in weeks since initial diagnosis).
    - Similarly, the start and end of the events should not be posterior to the overall survival, also measured in weeks. If it is, this is flagged as inconsistent. In addition to treatments, this check also applies to response and surgery events.
    - Analogously, if the start of the treatment (surgery, radiation, targeted or pharmacotherapy) or the start of the response to therapy is before the diagnosis, a warning is raised. This is also the case when such events are indicated to happen in the future. The same check is done for the end of therapies (radiation, targeted or pharmacotherapy).
    - Validations also check for cases with incomplete follow-ups. These cases would be those in which death by colon cancer is indicated in the vital status, but the response value is marked as complete response. For such cases, the start of the last response and the overall survival are also validated.
    - The duration of the therapy is checked and flagged if a non-surgery therapy started and ended in the same week as initial diagnosis or if the duration (difference between its start and end) is less than one week—since these treatments normally last longer. Analogously, negative duration of treatments are also flagged.
    - Detection of new series of treatments without indicating recurrence is also part of the validations.
- Inconsistent values
    - Suspicious values in the description of pharmacotherapy. If the pharmacotherapy scheme is set as "Other" for a given patient, there must be a pharmacotherapy scheme description. If not, a warning is created. When it is provided, the value of this field is checked and flagged if it consists of non-descriptive values (e.g., "unknown" or "NULL") or if the substances used were not specified.
    - Inconsistencies within surgical data: surgery location and surgery type. List of normal and possible pairs is shown in Table 2; possible pairs trigger a warning, all other combinations (considered impossible) trigger an error.
    - Mismatch between histopathological and surgical data, when a location of tumor is substantially different from surgery location.
    - Mapping of pTNM values to UICC stage based on the UICC standard. Consistency between provided values of the UICC stage and the pTNM categories is checked for those samples in which the UICC stage and the three categories are provided. A new UICC stage is computed using the values of the three pTNM categories, taking into account the UICC version. The resulting stage is compared with

Surgery location (ICD-10 code)	Consistent surgery type	Suspicious surgery type
C18.0 C18.1 C18.2	"Right hemicolectomy"	"Pan-procto colectomy", "Total colectomy"
C18.3	"Right hemicolectomy"	"Pan-procto colectomy", "Total colectomy", "Transverse colectomy"
C18.4	"Transverse colectomy"	"Left hemicolectomy", "Pan-procto colectomy", "Right hemicolectomy", "Total colectomy"
C18.5	"Left hemicolectomy"	"Abdomino-perineal resection", "Pan-procto colectomy", "Total colectomy", "Sigmoid colectomy", "Transverse colectomy"
C18.6	"Left hemicolectomy"	"Abdomino-perineal resection", "Pan-procto colectomy", "Total colectomy", "Sigmoid colectomy
C18.7	"Sigmoid colectomy", "Left hemicolectomy"	"Abdomino-perineal resection", "Low anterior colon resection", "Pan-procto colectomy", "Total colectomy"
C18.9*	"Left hemicolectomy", "Pan-procto colectomy", "Right hemicolectomy", "Sigmoid colectomy", "Total colectomy", "Transverse colectomy"	"Abdomino-perineal resection", "Low anterior colon resection"
C19, C19.9	"Anterior resection of rectum", "Endo-rectal tumor resection", "Low anterior colon resection", "Sigmoid colectomy"	"Abdomino-perineal resection", "Left hemicolectomy", "Pan-procto colectomy", "Total colectomy"
C20, C20.9	"Abdomino-perineal resection", "Anterior resection of rectum", "Endo-rectal tumor resection", "Low anterior colon resection"	"Left hemicolectomy", "Pan-procto colectomy", "Total colectomy"

Table 2. Surgery consistancy.

the provided one, raising a warning if they do not match. A warning is also created if the UICC stage can not be computed—for instance, due to the incompatibility of the pTNM values and the given UICC version, or the combination of certain pTNM values with the UICC stage (e.g., pNX combined with defined stage, where it is not clear on what basis the stage was determined).

- Warnings on missing variables
  - If the UICC stage is provided and is different from stage IV (i.e., stage is lower than IV and hence M0), it indicates that the value of the N category should be valid (not Nx). Otherwise, a warning is set. Note: The model does not accept uncertain determination of stage, such as "at least stage I or II". Therefore N has to be defined and different from Nx when M0.
  - Patients with a known vital status must have the timestamp for this value, otherwise a warning is created.
  - A warning is also created if the values below are missing:
    - Response to therapy
    - Start of response to therapy

Practical experiences. It was quite surprising to see to what high levels data was incomplete and inconsistent in some cases when arrived from the source health care organizations to which the biobanks are connected. Typical issues were inconsistent survival information (e.g., treatments provided even after patient died), mismatching pTNM values with UICC stage values (which is a relatively simple check based on UICC standard), mismatch between location of tumor and type/location of surgery, and various other problems related to outcomes of treatment (like starting a new chemotherapy after complete response to therapy, without indicating any recurrence). These problems required biobanks to go back to the source health care organization in order to retrieve and validate ground truth and even led to some cases being removed and replaced by others during the collection process, as sufficiently reliable data could not be retrieved retrospectively. Some other data quality issues became obvious only after pooling larger amounts of data from multiple biobanks together, so that substantial differences in distributions became more visible (e.g., analyzing overall survival data with respect to the UICC stage).

Usage Notes

Data access. There are no geographic restrictions on access to the cohort; applications are evaluated independently of the applicant's country of origin. Both academic and industrial researchers are eligible to request access, provided that their proposed projects comply with the scientific purpose and ethical framework of the cohort. All approved users receive access to the data free of charge, ensuring that the resource remains broadly available to the international research community. In order to streamline the access process and maximize the benefits of BBMRI-ERIC acting as a data deposition facility and data controller for the CRC-Cohort Dataset, BBMRI-ERIC developed the access procedure for centralised datasets (https://www.bbmri-eric.eu/services/access-policies/). This is an extension to the generic BBMRI-ERIC Access Policy (https://www.bbmri-eric.eu/wp-content/uploads/AoM_10_8_Access-Policy_FINAL_EU.pdf). The procedure for centralised datasets allows

for implementation of the committee-controlled access mechanism in the federated environment, but also allows the BBMRI-ERIC to reach the access decision within one month.

Access to the centrally stored data is provided directly by BBMRI-ERIC. Based on CRC-Cohort Dataset DPP, access can be provided to either the original pseudonymized data (subject to data minimization for a particular purpose/project) or to de-facto anonymized data (subject to data quality loss). Either way, the data can be only accessed after signing a "Data Transfer Agreement" with BBMRI-ERIC, which, among other things, defines the scope of the activities involving the data.

The Access Procedure for BBMRI-ERIC centralised datasets. comprises the following 5 steps (the description is simplified).

Step-1 Registration of the requester.] BBMRI-ERIC verifies the identity of each requester and his/her institutional affiliation (employee status).

Step-2 Request of samples/data.] A requester files a request for access to samples/data via the BBMRI-ERIC Negotiator (https://negotiator.bbmri-eric.eu/login.xhtml) in response to which the provider responds with Availability Information. Both the request and Availability Information are treated confidentially by all parties involved.

Step-3 Access control & delivery of samples/data.] After receiving adequate Availability Information, the requester follows up directly with the provider (BBMRI-ERIC) to provide any additional information needed to assess whether access can be granted and to perform any additional required steps—e.g., signing Material/Data Transfer Agreement (DTA) (see Annex E of DPP). Material/Data is transferred directly between provider and requester. *As BBMRI-ERIC is the data controller, special provisions have to be defined and agreed on a case by case basis and made available to the requesters*. For controlling access to the dataset, BBMRI-ERIC has established an Access Committee to ensure that the due project approvals are in place, particularly when releasing pseudonymous (personal) data.

The Access Committee is comprised of 3 experts: a designated Data Manager of the CRC-Cohort Dataset, one appointed person from Common Service ELSI for the ethics check, and one medical expert nominated by BBMRI-ERIC.

Operations of the Access Committee and involvement of contributing partner biobanks. the Access Committee is expected to deliver the decision within 1 month of a request being submitted, based on the following procedure.

Step-3a After receiving an access request (project proposal), the Data Manager checks whether the access request conforms with the formal requirements: (a) the identity of the requester is known and their institutional affiliation is provided, (b) the request contains the project description.

Step-3b If the formal requirements are fulfilled, the medical expert assesses relevance of the project to the scope of the CRC-Cohort Dataset. The cohort should enable a large spectrum of different types of research and is, therefore, not designed for or restricted to specific research questions. Thus, this assessment does not evaluate the scientific novelty, priority, nor impact, as long as the research is related to colorectal cancer (and has ethical endorsement, see Step-3d). Specifically, the assessment verifies that the proposed analyses are methodologically coherent with the available data, and reviews if the requested data is adequately minimized for the particular research purpose, which is requirement of GDPR. In case data is not minimized in the request, data minimization is negotiated and confirmed with the applicant.

Step-3c If the project is within the scope, the Access Committee performs the Ethics Check procedure (either Expedited or Full, depending on whether a sufficient previous ethics vote has been provided, see Annex G of DPP for more details).

Step-3d If all the previous steps conclude successfully, the Access Committee contacts all the contributing partner biobanks for a possible veto of the release. If only a subset of the CRC-Cohort Dataset is requested, only the partner biobanks whose data is being requested are contacted. BBMRI-ERIC serves as the data controller; however, as the biobanks are the contributing entities that make the cohort possible, they retain sovereignty over the subset of data they have provided. Accordingly, each biobank may veto only the release of the data it has contributed, and solely on predefined grounds related to legal, ethical or consent-related constraints. Vetoes based on scientific topic, novelty or perceived merit of the research are explicitly excluded.

To ensure a standardised, swift and transparent process, biobanks are given 14 days to submit a possible veto. If no response is received within this period, approval is assumed and the access procedure continues without delay. This time-limited veto mechanism ensures that biobank sovereignty is respected while preventing unresponsiveness from obstructing justified data access.

Step-3e After the access is approved by the Access Committee, BBMRI-ERIC signs a DTA with the requester for the use of the data for the specific project.

Step-4 Return of results. Providers need to collect reports on project outcomes for accountability purposes regarding the utilization of the BBMRI-ERIC infrastructure. Providers are encouraged to require the return of derived data from the requester and integrate this requirement into the Material/Data Transfer Agreement.

Step-5 Request completion notification. For each request obtained via BBMRI-ERIC, for which Availability Information is provided according to Step-2 and where Step-3 is completed, the provider is required to inform BBMRI-ERIC whether the request is completed successfully or whether it failed. In case a request fails, reasons for failure have to be specified. For successfully completed requests, the provider will report project outcomes to BBMRI-ERIC.

Thanks to this time-limited veto procedure, 1-month overall time limit for decision making is possible—if the contributed biobank does not react, it is taken as an approval. Note that the DTA also transfers liability on the requester to ensure due data security measures are applied when processing the data.

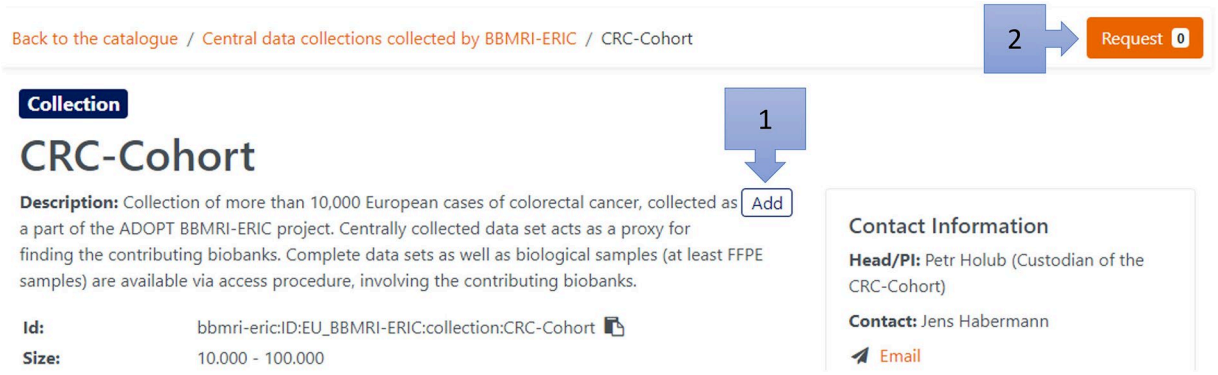

Fig. 3 Example of selecting CRC-Cohort Dataset using BBMRI-ERIC Directory.

Example usage. The CRC-Cohort Dataset has enabled existing, well-established biobanks in Europe to connect with BBMRI-ERIC to create a comprehensive cohort for CRC research, which includes participants from several European countries, facilitating access to this data for future research use. The primary scope of the CRC-Cohort Dataset was to test the feasibility of integrating data on CRC from different countries and making them accessible through a common access mechanism as a part of BBMRI-ERIC. Stage II CRC is an example of a specific medical need that can be addressed with such a cohort, as there is still lacking evidence for optimum treatment of the patients, and whether surgery is sufficient or chemotherapy is beneficial to prevent recurrence of the disease for a particular patient. Another more specific need is accessing large datasets from different countries for training of machine learning algorithms for digital pathology.

Specific practical goals of the CRC-Cohort Dataset are to support research in the following areas:

- testing the feasibility of retrospectively integrating into a research cohort biological samples and related medical data generated in the context of health services;
- identification of biomarkers effective for predicting prognosis and selecting therapy for patients with stage II CRC;
- providing a collection of digitized images of the histopathological sections together with outcome data for the development of so called imaging biomarkers by machine learning methods;
- establishing a benchmark dataset for evaluation of quality of anonymization techniques and related residual risk of re-identification by BBMRI-ERIC.

Existence of the centrally collected data supports researchers in formulating medically relevant projects and improving study designs.

Practical instructions. The CRC-Cohort Dataset can be discovered via BBMRI-ERIC Directory (a direct link to access the collection at the time of writing this paper is https://directory.bbmri-eric.eu/ERIC/directory/#/collection/bbmri-eric:ID:EU_BBMRI-ERIC:collection:CRC-Cohort?search=crc) and/or the Federated Platform. From there, it can be added to the access request in the BBMRI-ERIC Negotiator. A practical example how to select CRC-Cohort Dataset in the Directory is shown in Fig. 3. Data requesters can ask for access to the study in the cBioPortal instance, which allows statistical analyses of clinical and genomic data and WSI visualisation through xOpat. BBMRI-ERIC Negotiator also brings the possibility of requesting access to individual level clinical, genomic and/or WSI pseudonymised data, as well as to the associated biosamples that are part of the collections of the participating biobanks.

Secure data transfers. A secure mechanism is available for bulk network transfer of CRC-Cohort Dataset data, including WSI of slides in OME-TIFF format. The approved requester must provide access to a storage space that is accessible via Internet, requires user authentication, and has sufficient capacity to contain the requested data. Several common storage options are supported, including cloud object storage services, SFTP and GridFTP. In addition, the requester will need to provide a *public* cryptographic encryption key from a public-private key pair, which will be used to protect the shipped data with an envelope encryption scheme—currently implemented with the Crypt4GH tool[38]; upon reception, the requester will be able to decrypt the data using his/her private key. The scheme ensures that the data is shipped always protected by two factors—i.e., the envelope encryption plus either the encrypted transmission channel while in flight, or the storage authentication while at rest at the destination.

Data availability

The CRC-Cohort Dataset[29] is available at https://directory.bbmri-eric.eu/ERIC/directory/#/collection/bbmri-eric:ID:EU_BBMRI-ERIC:collection:CRC-Cohort?search=crc

Code availability

- Availability of the data quality framework.
  - Implementation of XSDs for simple conformance and completeness validation available at https://doi.org/10.5281/zenodo.7514312[17] (as stated above, there are XSDs available for strict and partial validation).

– Implementation of advanced quality checks in R is available at https://github.com/BBMRI-ERIC/CRC-Cohort-QC.
- openEHR-based data mapping is available from https://github.com/crs4/crc_cohort_modelling/.
- openEHR-based data mapping tooling for FHIR is available from https://github.com/samply/bbmri-fhir-gen.

Received: 27 January 2025; Accepted: 5 February 2026;
Published online: 12 March 2026

References

1. Morgan, E. *et al*. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from globocan. *Gut* **72**, 338–344, https://doi.org/10.1136/gutjnl-2022-327736 (2023).
2. Sawicki, T. *et al*. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers* **13**, 2025 (2021).
3. Baxter, N. N. *et al*. Adjuvant therapy for stage ii colon cancer: Asco guideline update. *Journal of Clinical Oncology* **40**, 892–910, https://doi.org/10.1200/JCO.21.02538 (2022).
4. Shaukat, A. & Levin, T. R. Current and future colorectal cancer screening strategies. *Nature Reviews Gastroenterology & Hepatology* **19**, 521–531 (2022).
5. Bretthauer, M. Colorectal cancer screening. *Journal of internal medicine* **270**, 87–98 (2011).
6. Bailey, J. R., Aggarwal, A. & Imperiale, T. F. Colorectal cancer screening: stool dna and other noninvasive modalities. *Gut and Liver* **10**, 204 (2016).
7. Harlé, A. Cell-free dna in the management of colorectal cancer. *Tumor Liquid Biopsies* 253–261 (2020).
8. Sepulveda, A. R. *et al*. Molecular Biomarkers for the Evaluation of Colorectal Cancer: Guideline From the American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and American Society of Clinical Oncology. *Archives of Pathology & Laboratory Medicine* **141**, 625–657, https://doi.org/10.5858/arpa.2016-0554-CP (2017).
9. Tan, C. & Du, X. KRAS mutation testing in metastatic colorectal cancer. *World journal of gastroenterology: WJG* **18**, 5171 (2012).
10. Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D. & Hitchins, M. P. Milestones of lynch syndrome: 1895–2015. *Nature Reviews Cancer* **15**, 181–194 (2015).
11. Ho, C. *et al*. A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer. *Scientific Reports* **12**, 1–9 (2022).
12. Yu, G. *et al*. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. *Nature communications* **12**, 1–13 (2021).
13. Wulczyn, E. *et al*. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ digital medicine* **4**, 71 (2021).
14. Vacante, M., Borzì, A. M., Basile, F. & Biondi, A. Biomarkers in colorectal cancer: Current clinical utility and future perspectives. *World journal of clinical cases* **6**, 869 (2018).
15. Nguyen, L. H., Goel, A. & Chung, D. C. Pathways of colorectal carcinogenesis. *Gastroenterology* **158**, 291–302 (2020).
16. Muscholl, M., Lablans, M., Wagner, T. O. & Ückert, F. OSSE–open source registry software solution. *Orphanet journal of rare diseases* **9**, O9 (2014).
17. Proynova, R., Ataian, M., Törnwall, O. & Holub, P. Technical documentation for contributing to BBMRI- ERIC CRC-Cohort, https://doi.org/10.5281/zenodo.7514312 (2023).
18. Mate, S. *et al*. Pan-european data harmonization for biobanks in ADOPT BBMRI-ERIC. *Applied Clinical Informatics* **10**, 679–692, https://doi.org/10.1055/s-0039-1695793 (2019).
19. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians* **73**, 17–48, https://doi.org/10.3322/caac.21763 (2023).
20. WHO report on cancer: setting priorities, investing wisely and providing care for all (World Health Organization, 2020).
21. Howlader, N. *et al*. Tech. Rep., National Cancer Institute, Bethesda, MD (2020). Based on November 2019 SEER data submission, posted to the SEER web site.
22. Nunes, L. *et al*. Prognostic genome and transcriptome signatures in colorectal cancers. *Nature* **633**, 137–146, https://doi.org/10.1038/s41586-024-07769-3 (2024).
23. Cerami, E. *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401–404, https://doi.org/10.1158/2159-8290.CD-12-0095 (2012).
24. Gao, J. *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling* **6**, pl1–pl1 (2013).
25. de Bruijn, I. *et al*. Analysis and visualization of longitudinal genomic and clinical data from the aacr project genie biopharma collaborative in cbioportal. *Cancer research* **83**, 3861–3867 (2023).
26. xopat: explainable open pathology analysis tool. In Computer Graphics Forum, vol. 42, 63-73 (Wiley Online Library, 2023).
27. Atalag, K. *et al*. openehr—a semantically enabled, vendor-independent health computing platform. *University College London*, (2016).
28. Principles of health interoperability: SNOMED CT, HL7 and FHIR (Springer, 2016).
29. BBMRI-ERIC. CRC-Cohort. https://directory.bbmri-eric.eu/ERIC/directory/#/collection/bbmri-eric:ID:EU_BBMRI-ERIC:collection:CRC-Cohort.
30. TNM classification of malignant tumours (John Wiley & Sons, 671 2017).
31. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics* **4**, 27 (2013).
32. Linkert, M. *et al*. Metadata matters: access to image data in the real world. *Journal of Cell Biology* **189**, 777–782 (2010).
33. Leigh, R. *et al*. OME files-an open source reference library for the OME-XML metadata model and the OME-TIFF file format. *BioRxiv* 088740 (2017).
34. Besson, S. *et al*. Bringing open data to whole slide imaging. In *European Congress on Digital Pathology*, 3-10 (Springer, 2019).
35. Schaaf, J. *et al*. OSSE goes FAIR-implementation of the FAIR data principles for an open-source registry for rare diseases. *Studies in health technology and informatics* **253**, 209–213 (2018).
36. Kadioglu, D. *et al*. Samply.MDR - a metadata repository and its application in various research networks. *Studies in Health Technology and Informatics* **253**, 50–54, https://doi.org/10.3233/978-1-61499-896-9-50 (2018).
37. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* **5**, 299–314 (1996).
38. Senf, A. *et al*. Crypt4GH: a file format standard enabling native access to encrypted data. *Bioinformatics* **37**, 2753–2754, https://doi.org/10.1093/bioinformatics/btab087 (2021).

Acknowledgements

This work has been co/funded by ADOPT BBMRI-ERIC project supported by EU Horizon 2020, grant agreement no. 676550; EOSC-Life project, supported by EU Horizon 2020, grant agreement no. 824087, as a part of WP1 Demonstrators under APPID 1228 "Cloudification of BBMRI-ERIC CRC-Cohort and its Digital Pathology Imaging"; the HealthData@EU Pilot project, funded under the EU4Health Programme, grant agreement no.

101079839, and the XDATA Project, financed by the Sardinian Regional Government. Parts of this work have received funding from the Austrian Science Fund (FWF), Austria, Project P-32554 (Explainable Artificial Intelligence). WSI visualization infrastructructure has been partially supported by the BioMedAI TWINNING project, supported by EU Horizon Europe, grant agreement no. 101079183. In addition, this work has been supported by funding of BBMRI-ERIC National and Organizational Nodes: Czech Ministry of Education, Youth and Sports (LM2018125–BBMRI-CZ); the Austrian Federal Ministry for Education, Science and Research (BMBWF-10.470/0010-V3c/2018; BBMRI.at).

Author contributions

P.H. was IT coordinator of the collection, designed and co-implemented data quality mechanisms and visualizations, led development of Data Protection Policy, designed final reimbursement model, led the EOSC-Life WP1 Demonstrator and led editing the paper. R.P. and M.A. implemented the CCDC system. F.S. co-implemented data quality mechanisms. D.A. and F.Ü. and O.V. contributed to the development of data model. K.Z. co-led design of the cohort, led development of histopathological imaging part of the cohort, contributed to the development of the data model. M.L. co-led design of the cohort and organized recruitment of source biobanks in collaboration with BBMRI-ERIC National and Organizational Nodes. M.H. contributed to the design of the cohort and data model. O.C., G.M., D.V., B.P., A.H., M.B., G.V., E.C. contributed to the development of the data model. O.T. managed the contracting with source biobanks and implemented reimbursement model, contributed to development of Data Protection Policy. A.L.K. managed the contracting with source biobanks and implemented reimbursement model. E.S. co-designed final reimbursement model. I.S. co-led to development of Data Protection Policy. V.H. contributed to biobanks support in data provisioning. C.A. contributed to the documentation of the cohort. S.M., H.-U.P developed tooling for data harmonization. L.M., M.M., A.F. implemented data hosting infrastructure. E.G.A. contributed to cohort data management and integration into Federated Platform. J.H. integrated the cohort and the xOpat viewer into cBioPortal. F.F., C.M., A.S., G.D., M.D.R., V.M. implemented openEHR data transformation and hosting and co-developed OMOP and HL7 FHIR mapping. H.M. co-developed OMOP and HL7 FHIR mapping and contributed to data provisioning. R.R. contributed to data provisioning. L.P., M.K., M.E.P. implemented OME-TIFF WSI data storage. S.L. contributed automated Workflow RO-Crate generation. A.K., K.S., A.R.A., B.H., C.F., D.P.B., F.A.G., F.V., G.C.C., I.Z., M.D.B., J.V., K.K., M.P.M., S.M., T.B-H., T.B., T.S., T.S., V.P., Y.B. implemented the data preparation process at contributing institutions and provided the data into the CRC-Cohort. All authors reviewed the manuscript.

Competing interests

K.Z. is founder and CEO of Zatloukal Innovations GmbH.

Additional information

Correspondence and requests for materials should be addressed to P.H. or E.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Petr Holub[1,2]✉, Outi Törnwall[1], Eva Garcia Alvarez[1]✉, Rumyana Proynova[3], Florian Stampe[3], Saher Maqsood[3], Maxmilian Ataian[4], Irene Schlünder[1], Olli Carpén[5], Gerrit Meijer[6], Rudolf Nenutil[7], Dalibor Valík[8,9], Barbara Parodi[10], Annemieke Hiemstra[6], Mariska Bierkens[6], Esmeralda Castanos-Vélez[11], Frank Ückert[4], Diogo Alexandre[3], Ondřej Vojtíšek[7], Anna-Liisa Bader[1], Birgit Simell[1], Caitlin Ahern[1], Vitomir Horvat[1], Erik Steinfelder[1], Matteo Gnocchi[12], Marco Moscatelli[12], Alexander Fürbaß[1], Jiří Horák[1], Francesca Frexia[13], Cecilia Mascia[13], Alessandro Sulis[13], Giovanni Delussu[13], Mauro Del Rio[13], Vittorio Meloni[13], Luca Pireddu[13], Simone Leo[13], Marco Enrico Piras[13], Martin Kačenga[2], Stéphanie Gofflot[14], Sebastian Mate[15], Hans-Ulrich Prokosch[15,16], Paolo Romano[10], Daniela Pistillo[17], Michael Hoffmeister[18], Alexander Brobeil[19], Amila Kugic[20], Berthold Huppertz[21], Valentina Paleari[17], Heimo Müller[22], Robert Reihs[22], Timo Gemoll[23],

Yannick Bantel[24], Tobias Sjöblom[25], Kyriacos Kyriacou[26], Simona di Martino[27], Gennaro Ciliberto[28], Ann-Kristin Kock-Schoppenhauer[29], Martina Oberländer[30], Jens K. Habermann[1,29], Gabriele Husmann[31], Per-Henrik D. Edqvist[25], Inti Zlobec[32], Martin D. Berger[33], Lars Boeckmann[23,34], Fabienne George[35], Tom Southerington[36], Daniel P. Brucker[37], Laurence Faugeras[35], Joanna Vella[38], Alex Felice[38], Malcolm Pace[38], Chiara Fallerini[39], Alessandra Renieri[39], Andreas Hadjisavvas[40], Karine Sargsyan[22,41,42], Maria A. Loizidou[26], Tatiana Besse-Hammer[43], Franziska Vogl[20], Jan-Eric Litton[44], Michael Hummel[11], Kurt Zatloukal[45] & Marialuisa Lavitrano[46]

[1]BBMRI-ERIC, Neue Stiftingtalstrasse 2/B/6, 8010, Graz, Austria. [2]Insitute of Computer Science, Masaryk University, Šumavská 525/33, 60200, Brno, Czech Republic. [3]BBMRI.de/German Biobank Alliance and German Cancer Research Center, Heidelberg, Germany. [4]Institute of Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [5]University of Helsinki, Helsinki, Finland. [6]Netherlands Cancer Institute (NKI), Amsterdam, The Netherlands. [7]Masaryk Memorial Cancer Institute, Žlutý kopec 543/7, 60200, Brno, Czech Republic. [8]Faculty of Medicine, Masaryk University, Kamenice 5, 62700, Brno, Czech Republic. [9]Faculty Hospital Brno, Jihlavská 340/20, 62500, Brno, Czech Republic. [10]IRCCS Ospedale Policlinico San Martino, Largo Rosanna Benzi 10, 16132, Genova, Italy. [11]Charité - Universitätsmedizin Berlin, Berlin, Germany. [12]National Research Council—Institute of Biomedical Technologies, Segrate, Italy. [13]CRS4, Visual and Data-Intensive Computing Group, Loc. Piscina Manna, 09050, Pula, CA, Italy. [14]Biotheque Hospitalo-Universitaire de Liege, CHU de Liege, Avenue Hippocrate B23+1, 4000, Liege, Belgium. [15]Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Krankenhausstraße 12, 91054, Erlangen, Germany. [16]Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Wetterkreuz 15, 91058, Erlangen-Tennenlohe, Germany. [17]Center for Biological Resources, IRCCS Humanitas Research Hospital, Via Manzoni 56, 20089, Rozzano, MI, Italy. [18]German Cancer Research Center (DKFZ), Division of Clinical Epidemiology of Early Cancer Detection, Im Neuenheimer Feld 581, 69120, Heidelberg, Germany. [19]Institute of Pathology; Tissue Bank of the National Center for Tumor Diseases (NCT), University of Heidelberg, Heidelberg, Germany. [20]Biobank Graz, Medical University of Graz, Neue Stiftingtalstraße 2/B/2, 8010, Graz, Austria. [21]Medical University of Graz, Division of Cell Biology, Histology and Embryology, Gottfried Schatz Research Center, Graz, Austria. [22]International Biobanking and Education, Medical University Graz, Neue Stiftingtalstrasse 2, 8010, Graz, Austria. [23]Section for Translational Surgical Oncology and Biobanking, Department of Surgery, University of Luebeck and University Hospital Clinic Schleswig-Holstein, Ratzeburger Allee 160, 23538, Lübeck, Germany. [24]CeGaT GmbH, Paul-Ehrlich-Straße 23, D-72076, Tübingen, Germany. [25]Uppsala University, Dept Immunology, Genetics and Pathology, Dag Hammarskjöldsväg 20, 75185, Uppsala, Sweden. [26]The Cyprus Institute of Neurology and Genetics, Iroon Avenue, Agios Dometios, 2371, Cyprus. [27]Pathology Unit and Biobank IRCCS- Regina Elena National Cancer Institute, via elio chianesi 53, Rome, Italy. [28]IRCCS National Cancer institute Regina Elena, Via Elio Chianesi 53, 00144, Rome, Italy. [29]IT Center for Clinical Research, University of Lübeck and University Hospital Schleswig-Holstein, Lübeck, Germany. [30]Interdisciplinary Biobank Centre - Lübeck, University of Luebeck and University Hospital Clinic Schleswig-Holstein, Ratzeburger Allee 160, 23538, Lübeck, Germany. [31]University Cancer Center Frankfurt, University Hospital Frankfurt, Theodor-Stern-Kai 7, 60590, Frankfurt, Germany. [32]Institute of Tissue Medicine and Pathology, University of Bern, Murtenstrasse 31, CH-3008, Bern, Switzerland. [33]Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland. [34]Clinic and Policlinic for Dermatology and Venereology, University Medical Center Rostock, Strempelstraße 13, 18057, Rostock, Germany. [35]CHU UCL Namur 1, av Gaston Therasse, 5530, Yvoir, Belgium. [36]University of Turku, 20014, Turun yliopisto, Turku, Finland. [37]University Hospital Frankfurt, 60590, Frankfurt, Germany. [38]The Centre for Molecular Medicine and Biobanking, University of Malta, MSD2080, Msida, Malta. [39]University of Siena, Viale Bracci, 1, 53100, Siena, Italy. [40]Department of Cancer Genetics, Therapeutics & Ultrastructural Pathology The Cyprus, Iroon Avenue, 2371, Ayios Dometios, Nicosia, Cyprus. [41]Department of Medical genetics, Yerevan State Medical University Yerevan, Yerevan, Armenia. [42]Cancer Center, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [43]CHU Brugmann, Place A. Van Gehuchten 4, 1020, Bruxelles, Belgium. [44]Karolinska Instituet, Stockholm, Sweden. [45]Diagnostic and Research Center for Molecular Biomedicine, Medical University Graz, Neue Stiftingtalstrasse 6, 8010, Graz, Austria. [46]University Milano Bicocca, Milan, Italy. ✉e-mail: petr.holub@bbmri-eric.eu; eva.garcia-alvarez@bbmri-eric.eu