BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort): Data Protection Policy

Petr Holub, Irene Schlünder, Kurt Zatloukal, Outi Törnwall, Ayodeji Adeniran, Marialuisa Lavitrano, Michael Hummel

Version: 1.5, 2023-05-12
DOI: 10.5281/zenodo.7930536

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550.

Contents

Summary

Colorectal Cancer Cohort (CRC-Cohort) is developed by BBMRI-ERIC, its BBMRI-ERIC National/Organisational Nodes, and BBMRI-ERIC partner biobanks. The CRC-Cohort is developed as a part of the ADOPT BBMRI-ERIC project[1] and will become permanent asset of BBMRI-ERIC research infrastructure after the end of the project in order to enable research to improve treatment of the colorectal cancer. The data collection should provide broad European coverage and sufficient number of research participants in order to enable research that was impossible before. The data collection process runs within the ADOPT BBMRI-ERIC project with the target 10,000 cases are expected to be collected until end of March 2018. The juridical person liable for establishment and operations of the cohort is BBMRI-ERIC. Pseudonymized data will be collected from individual partner biobanks based on a common data model developed and agreed upon within the ADOPT BBMRI-ERIC project. BBMRI-ERIC will ensure adequate data protection of the centrally collected data. This document describes data protection measure implemented and is intended for the data protection authorities supervising contributing partner biobanks and data protection experts in ethics committees.

The document is structured as follows. Section 1 summarizes objectives of the CRC-Cohort, its legal framework and basic organizational aspects. Section 2 describes data collection and integration process, together with measures for quality checking and assurance. Access modes for the data set are discussed in Section 3. Overview of tools on which implementation of the CRC-Cohort relies is provided in Section 4. More in-depth discussion of combination of technical and organizational measures to ensure data security and protect privacy of the persons contributing their data to CRC-Cohort is provided in Section 5. Risk analysis based on STRIDE and LINDDUN is available in Section 6. History of the document is available in Appendix A. Other relevant BBMRI-ERIC documents, and particularly those that are not (yet) publicly available, are provided as appendices.

[1] This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550 (implementAtion anD OPeration of the gateway for healTh into BBMRI-ERIC – ADOPT BBMRI-ERIC).

1. Introduction

This document describes design and practical implementation of Colorectal Cancer Cohort (CRC-Cohort) developed within ADOPT BBMRI-ERIC project,[2] with particular attention to the data protection aspect of the cohort.

1.1. Purpose of the cohort

BBMRI-ERIC,[3] an European research infrastructure of biobanks and biomolecular resources, missions to facilitate access to the biobanked samples and data in Europe. This has become boosted through an EU-funded (H2020) project, ADOPT BBMRI-ERIC project, where the access to European biobanks is piloted through a colorectal cancer cohort use case established by gathering of existing colon cancer data sets from different biobanks in Europe. The aim is to enable the existing, well-established biobanks in Europe to connect with BBMRI-ERIC to provide data sets and, later on, after specific research projects have been submitted to and approved by biobanks, samples for future research use. The data sets are gathered, integrated, and made available centrally for the research community to query and identify their specific research questions in colorectal cancer. The access to the centrally stored data is provided by BBMRI-ERIC to researchers according to the access policy set out in Section 3 and Appendix I. Collecting, storing and granting access to the data is subject to data protection measures described in this document.

Particular practical goals of the CRC-Cohort:

- The cohort of *existing* 10,000 colorectal cancer cases with detailed pathological and clinical data and available tissue samples should demonstrate the feasibility of large scale collaboration within BBMRI-ERIC and generate a yet unprecedented resource for medical research. **The cohort should enable a large spectrum of different types of research and is, therefore, not designed for or restricted to a specific research questions.** However, some examples of the intended use are as follows:
- to identify biomarkers for predicting prognosis and selecting therapy for patients with stage II disease.
- to provide the digital images of the histopathological sections together with outcome data for the development of so called imaging biomarkers by machine learning.[4]
- to establish a benchmark data set for evaluation of quality of anonymization techniques and related residual risk of re-identification by BBMRI-ERIC.

[2] This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550 (implementAtion anD OPeration of the gateway for healTh into BBMRI-ERIC – ADOPT BBMRI-ERIC).

[3] Biobanking and Biomolecular resources Research Infrastructure-European Research Infrastructure Consortium, see `http://www.bbmri-eric.eu/`

[4] Organizations interested in this have to register to access anonymized data to develop and test their algorithms. Data can only be used for this purpose and may not be distributed further. The study is realized as a competition and the results should be jointly published in a scientific journal.

- to support researchers in formulating medically relevant projects and improve the study designs.

The procedures and IT tools developed within the ADOPT BBMRI-ERIC project project and particularly CRC-Cohort are expected to be reusable for similar future projects on different disease entities implemented using BBMRI-ERIC as an infrastructure. Thus more general goals are

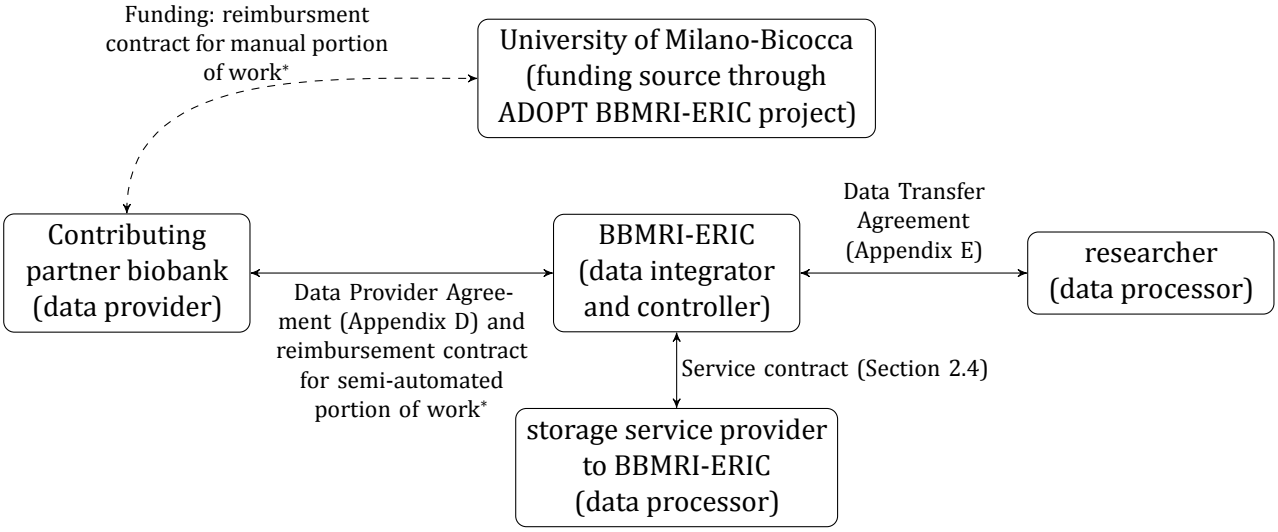
- to accelerate the future pan-European studies based on biobank data and the disease specific patient electronic health record information on colon cancer and other diseases,
- to enable the connection between the European biobank information systems and the coded clinical IT systems,
- to demonstrate the benefits of operational distributed Research Infrastructure to advance high-quality research and innovation.

1.2. Legal setup

- *Juridical person, data controller and custodian:* BBMRI-ERIC, Austria.
- *Applicable jurisdiction for additional data or biological material:* jurisdiction applicable for the respective partner biobank.
- *Responsible person for the data collection:* Director General of BBMRI-ERIC (<dg@bbmri-eric.eu>)
- *Contact person for the cohort:* CRC-Cohort Manager (currently Fereniki Ioakeimidou[5])
- *Data protection officer:* Data Protection Officer of BBMRI-ERIC (<dpo@bbmri-eric.eu>)

[5] Fereniki Ioakeimidou, BBMRI-ERIC, Neue Stiftingtalstraße 2/B/6, 8045 Graz, Austria. Fax: +43 316 34 99 17-99. Phone: +43 316 34 99 170. Email: <contact@bbmri-eric.eu>.

Contracting model.



The reimbursement model works as follows:

- For biobanks contributing small number of cases, they are entitled to get reimbursed for manual-only work at 150 €/case by sub-contracted and reimbursed (via University of Milano-Bicocca).
- For the remaining biobanks, they will be reimbursed 75% semi-automated portion of work (via BBMRI-ERIC) and 25% manual portion of work (via University of Milano-Bicocca), resulting in 56.25 €/case.
- Those biobanks delivering before December 15, 2017 and within first 1820 cases, are entitled to bonus of 25 €/case, implemented by increasing proportion of reimbursed manual portion of work (55% semi-automated portion of work and 45% manual portion of work).
- [Referring to * in the diagram.] Implementation of reimbursement from ADOPT BBMRI-ERIC project depends on whether the institution hosting the biobanks is a partner of ADOPT BBMRI-ERIC project or not. In former case, the reimbursement will be handled by transferring corresponding budget to the partner.[6] For the latter case, the biobank will be sub-contracted and reimbursed.

Preliminary General Data Protection Regulation (GDPR) Compliance. CRC-Cohort collection process and access procedure is designed to be compatible with the upcoming GDPR on

[6] Under the project financial regulations, an institution cannot be simultaneously a partner of the project and the sub-contractor eligible for reimbursement.

the European level.[7] It assumes operation under provisions of Article 89 §1 of GDPR for archival and research purposes. Depending on the practical interpretation of pseudonymization and anonymization under the GDPR and relevant Code(s) of Conduct (defined under Article 40 of GDPR), the pseudonymization and anonymization requirements may change. The pseudonymization and anonymization measures will be reviewed and adapted accordingly. This procedure is needed since interpretation of the GDPR and implementation on national level in different countries varies substantially at the time of writing this policy.

Compliance with IMI CoP and BBMRI-ERIC SPR This policy is intended to be compliant with the IMI Code of Practice on Secondary Use of Medical Data in Scientific Research Projects (see IMI CoP in Glossary). Pointers to relevant IMI CoP rules are given as notes throughout the document.

This policy is also designed to implement BBMRI-ERIC Security & Privacy Requirements [1].

1.3. Funding of the cohort

- Data collection process, 2015–2018: ADOPT BBMRI-ERIC project. This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550 (implementAtion anD OPeration of the gateway for healTh into BBMRI-ERIC – ADOPT BBMRI-ERIC).
- Cohort maintenance funding, 2018 and later: core budget of BBMRI-ERIC CS IT

1.4. Timeline

10/2015–09/2018 ADOPT BBMRI-ERIC project

10/2015 Definition of the biobank selection strategy

01/2016–06/2017 Biobank selection process

06/2017–03/2018 Data collection (manual or semi-automated)

03/2018 Milestone: 10,000 cases collected

03/2018 Start of the digital pathology competition (organized under ADOPT BBMRI-ERIC project Work Package 6)

09/2018 Start of access provisioning

[7] National implementation of the GDPR in the countries of the contributing partner biobanks is not yet known at the time of writing this document.

1.5. Relevant documents

- Statutes of BBMRI-ERIC [2]
- ADOPT Ethics Framework [3]
- Ethics Check (Appendix G)
- Template of Data Provider Agreement between a contributing partner biobank and BBMRI-ERIC (Appendix D)
- BBMRI-ERIC Access Policy (Appendix H) and Access Policy Amendment for CRC-Cohort (Appendix I)
- Terms and Conditions of Use of BBMRI-ERIC Services (Appendix F)
- Template of Data Transfer Agreement (DTA) between BBMRI-ERIC and researcher for accessing the data (Appendix E)

2. Data Collection and Storage

In total data of 10,000 cases of colorectal cancer cases shall be gathered in selected European biobanks that have formaline-fixed paraffin-embedded (FFPE) and/or fresh frozen tissue samples of surgical material available with associated clinical data (see inclusion criteria and data model in Appendix B). The availability of the biological material is part of the inclusion criteria, while only the data based on the common data model is collected and made available as a part of the CRC-Cohort.

A portion of ≈3,000 cases will be gathered manually. This means that the biobanks locate their existing cancer cases (clinical data linked with samples) and enter the required data set from the clinical files for each colorectal cancer case manually by using an secure web-based portal offered by BBMRI-ERIC—called Colorectal Cancer Data Collection software (further abbreviated as CCDC).[8]

The rest of the ≈7,000 cases are collected (semi-)automatically by using data transformation and possibly data-/text-mining tools in order to extract the data set associated with the samples from biobank information systems and hospital information systems of biobanks with already advanced IT systems. BBMRI-ERIC will provide tools and technical support to a limited extend for potential text-mining and work together with the biobanks' IT-staff to make the tools applicable for their use. After the data is extracted and transformed into the suitable format, it will be uploaded trough an API of the CCDC.

It is the responsibility of each contributing partner biobank to ensure that

- the contributed data has been collected and transferred to BBMRI-ERIC in compliance with all the relevant legal and ethical regulations,
- the contributed data is compatible with the purpose of the CRC-Cohort, its data protection policy and access procedures (defined in this document).

2.1. Informed consent and its withdrawal

It is the responsibility of each contributing partner biobank to ensure that they have informed consent from the patient to contribute the data to CRC-Cohort (either they already have it or they re-contact the patient), or that secondary use of the data was approved by the ethical review board to which the biobank adheres. It is the responsibility of the contributing partner biobank to maintain repository of all the relevant informed consents that entitle them to contribute the data to CRC-Cohort, or relevant decisions of the ethical review board granting secondary use of data for the CRC-Cohort.

Consent withdrawal is supported as follows: the patient contacts the contributing partner biobank, which notifies also BBMRI-ERIC (data manager of the cohort, see Section 1.2) that the data needs to be withdrawn from the cohort.

2.2. Process of providing data to BBMRI-ERIC

1. Data is prepared at the contributing partner biobank.

[8] We distinguish CCDC as a software tool developed to collect the data, and CRC-Cohort as the name of the cohort itself.

2. Patient non-speaking pseudonym must be generated by the contributing institution or by a trusted third party and delivered to the contributing institution. Pseudonyms must conform to the state-of-the-art pseudonymization guidelines (either pseudonyms independent of original identifiers, or keyed-hashing functions based pseudonyms with deletion of the key—see [4, Chapter 4]) and any national requirements on pseudonymization relevant for the contributing partner biobank. Initial pseudonymization must be only done by a person bound to medical or clinical secrecy.

3. Data is provided to BBMRI-ERIC via a secure web-based data collection application (for manual data entry – ≈3,000 cases) or via a secure REST upload API (for semi-automated or automated data upload – ≈7,000 cases)—both hosted at `https://ccdc.bbmri-eric.eu/`. The connection is secured using Transport Level Security (TLS) 1.1 or newer (newer version can be used depending on the client used) with X.509 server certificate is issued by commonly accepted root certificate authority (implemented using DigiCert[9] via GÉANT's Trusted Certificate Service[10] as of writing this policy).

4. Any data transformation tools or data-/text-mining tools will run inside the biobank, so that the biobank retains full control of the sensitive data prior to sending the pseudonymized data to BBMRI-ERIC as described in the next step.

5. Genetic data sufficient to single out an individual are not collected in the CRC-Cohort (only information on a few relevant mutations is collected, see Appendix B).

As a part of the data insertion, the constraints are checked based on the data model to ensure basic consistency of the inserted data.

2.3. Data integration

The following steps will be implemented as data integration into the central CRC-Cohort data set:

- Pseudonym integration: Each contributing partner biobank will be assigned a unique prefix to be prepended to the local pseudonyms generated by the biobank. This ensures global uniqueness of the identifier inside the CRC-Cohort.

- Both data collected via manual input via web interface of CCDC application and data provided using (semi-)automated processing will be stored in a SQL database with each patient record having identifier of the contributing partner biobank.

- Adding or modifying data in the CCDC is subject to the following authentication and authorization rules: *(a)* authentication is implemented using accounts managed centrally as a part of the CCDC application; *(b)* accounts are given strictly to individual physical persons and account sharing is forbidden by the policy, which has to be *a*

[9] `https://www.digicert.com/`
[10] `https://www.terena.org/activities/tcs/`

priori accepted by each contributing person; *(c)* each person is affiliated with one or more institutions; *(d)* authenticity is verified by good-entropy passwords; *(e)* each person can edit only the records contributed by the institution with which the person is affiliated; the exception is affiliates of BBMRI-ERIC who are entitled to access the whole data set for the purpose of the integration.

- As a part of the integration process, the data constraints will be reviewed and additional constraints are implemented if needed (Appendix B) in the CCDC software. The modified constraints are then implemented in the collection step as described above.
- The sanity of integrated data will be examined using statistical checks, to detect problems not detectable using data constraints on single data elements.

2.4. Data storage

BBMRI-ERIC will be the custodian and data controller of the integrated data set and will use service provider bound by service agreement (data processing contract) to store the data set (this service will be included as a part of BBMRI-ERIC Common Service IT to ensure long-term funding sustainability). The data storage will be in a SQL database. The following requirements will be imposed on any service provider of secure storage of personal data (naturally including pseudonymized data):

- Controlled physical access to the server hosting the virtual machine.
- Strict controlled access and regular evaluation of the security measures for the hypervisor hosting the virtual machines.
- All authorized personnel must treat the data confidentially, must observe security measures and must report any discovered problems that may result in compromising security.
- Control remote access over state-of-the-art secure protocols only.
- Network protection means (filtering and intrusion detection).
- Data is stored in encrypted storage with password for the encryption key provided at machine boot.
- Defined handling of security incidents in a manner transparent to BBMRI-ERIC.
- Data is backed up in a secure way (the same organizational controls apply for encryption and physical access to media).
- Data is provably removed/destroyed at the service provider after termination of the contract.

2.5. Quality control

Organizational measures and document rules.

- All the source code used for collecting the data for CRC-Cohort must be stored in a source code repository with versioning support.
- All configuration files must be stored in a repository with versioning support. Deployment of configuration files updates must be documented and the configuration files must include version identifier.
- All the documents generated regarding the CRC-Cohort need to be versioned, which each version specifying author(s), contributor(s), and reviewer(s), and description of changes. Independent of the technology used to generate the documentation, the technology must support marking line-level or word-level differences between revisions in order to ensure traceability.
- Unless restricted by privacy/security considerations, all the relevant documents shall be deposited in Zenodo,[11] in order to ensure assigning DOI identifiers and long-term preservation of the documents. This also holds for any released versions of individual documents.
- Privacy/security restricted documents will be stored as a part of BBMRI-ERIC Common Service IT in the infrastructure used for storing personal data.

Data quality.

- Each contributing partner biobanks must internally document how the data was collected/generated. Due to lack of common provenance model this can be documented in the common language—either English or the official language of the country the biobank resides (for the latter case the biobank should be able to provide reliable translation to English upon request). In particular, the following aspects must be recorded:
  - source system for the given data element,
  - whether the data was self-reported or measured,
  - the measurement method for the measured data,
  - any transformation applied to the data into the target format for the integration (for (semi-)automated processing this can be stored as versioned repository of processing tools, with clear tagging of the tool version and for adaptive tools also state of the trained system—for reproducibility reasons).

  This is compatible with international guidelines for meaningful data integration in the domain of cohorts [5, 6].

[11] `https://zenodo.org/`

3. Data Access Control

BBMRI-ERIC becomes a data controller and custodian of the collected data set. While ADOPT BBMRI-ERIC project has been used to collect the data set, the access procedure set up for the cohort is independent of the ADOPT BBMRI-ERIC project and will continue to exist as a part of core BBMRI-ERIC activities.

Access to the data set will be provided in three modes (shown also schematically in Figure 1, with more detailed component-based overview provided in Figure 3):

Mode-1 *Public (non-authenticated) search and browsing access to highly aggregated metadata describing the CRC-Cohort will be generated and published in the BBMRI-ERIC Directory*[12] *to ensure basic findability of the cohort.*

The statistics will be gathered from the CRC-Cohort pseudonymized data to fill in the metadata required by the BBMRI-ERIC Directory [7, 8] (see also Appendix C). Given the size of 10,000 or more cases in the CRC-Cohort and the flat data model, it can be considered highly aggregated statistical data with near zero risk of re-identification. Any search (querying) or browsing in the BBMRI-ERIC Directory is solely based on this data set.

Mode-2 *Secure authenticated search access by querying the anonymized data set with relatively low residual risk of re-identification (anonymization being k-anonymity with* $k \geq 5$ *considering all the data elements as pseudo-identifiers, or stronger) via search functionality using upcoming BBMRI-ERIC Locator/Finder*[13].

This access mode is designed to allow researchers assess more detailed preliminary availability of samples/data, to help them formulate relevant research projects. Secure authenticated access requires acceptance of conditions of use, which forbid any attempts to re-identify persons participating in CRC-Cohort; this is to mitigate residual risks after anonymization. See Section 5 for more discussion on anonymization.

Mode-3 *Secure authenticated access to the pseudonymized data (= personal data) controlled by Access Committee (see Appendix I) established for Mode-3 will include BBMRI-ERIC and biobank,*[14] *which provided the respective data considered to be released.* The pseudonyms are generated anew for each release of the data to avoid straightforward attacks based on reuse of the pseudonym. BBMRI-ERIC guarantees that the Access Committee decides on access within 1 month of submitting the request. Access is governed by Access Policy for CRC-Cohort (Appendix I).

The access can be provided only when all of the following conditions are met:

- The access to this data will be only provided to the researchers whose (a) identity have been verified, (b) who have research project(s) compatible with

[12] `https://directory.bbmri-eric.eu`

[13] BBMRI-ERIC Locator/Finder is scheduled for release in 2018.

[14] For efficient operations, multiple biobanks from the same country can be represented by a single representative, if acceptable in the given legal/organizational environment of the contributing biobank(s).

the purpose of the collected data or relevant legal exemptions, and (c) project has been successfully passed through the BBMRI-ERIC Ethics Check procedure[15] [3] and Appendix G.

- Data will be given based on data minimization principle,[16] i.e., access may be granted to anonymized data if it is sufficient to carrying out research, instead of pseudonymized data.
- Access (transfer of the data to the requester) is provided under Data Transfer Agreement (DTA) (Appendix E), which ensures that the data is processed under at least the same security requirements as this policy of BBMRI-ERIC and the data is used only for the given purpose.
- If transfer is requested to outside of European Union, it will be dealt with in compliance with the currently applicable data protection regulation (Directive 95/46/EC, later superseded by GDPR). In such a case the partner biobanks contributing to the released data set will be explicitly notified during the access approval procedure by Access Committee, to consider this circumstance when considering veto of the process (see access procedure defined in Appendix I).
- Consistently with the BBMRI-ERIC Access Policy [10], achieved results shall be communicated by the researcher back to BBMRI-ERIC, which further relays the information to the contributing partner biobanks to inform the research participants in compliance with their policies.
  The research data should be retained by the researcher long enough to allow reproducibility and verifiability of findings. Minimum retention period will be defined in contract between BBMRI-ERIC and the researcher (see Appendix E). The pseudonymized data and/or released anonymized data set must be deleted at termination of the DTA (Appendix E).
- Incidental findings achieved using pseudonymized data shall be reported by the researcher back to the contact person for the CRC-Cohort at BBMRI-ERIC, who establishes direct communication link between to researcher and the contributing partner biobanks in order to handle the incidental findings based on the policy of each contributing partner biobank.

The access in this mode can be requested via BBMRI-ERIC Negotiator[17] (either directly or via BBMRI-ERIC Directory or upcoming BBMRI-ERIC Locator/Finder), in compliance with the BBMRI-ERIC Harmonized Access Procedure [11] and Access Policy [10]. Before using the BBMRI-ERIC Negotiator, the requester must accept Terms and Conditions of Use of BBMRI-ERIC Services (Appendix F), which prevents the re-identification and privacy intrusion attempts.

Each requests consists of the following main parts:

[15] Note that for projects that have already been ethically assessed with positive outcome compatible to European regulations, the Ethics Check has a bypass mechanism called Expedited Ethics Check to avoid duplicate assessment.

[16] Rec.39; Art.5(1)(c) of GDPR [9].

[17] `https://negotiator.bbmri-eric.eu`

- identification of the researcher and their institutional affiliation,
- researchers consent to access policy of BBMRI-ERIC [10] and to conditions of use services (Appendix F – either already agreed to when reaching Negotiator from the Locator, or agreed to when ,
- identification and description of the project under which the data/material is requested,
- present existing ethics vote or if not available, apply for ad hoc ethics vote based on Ethics Check procedure (Appendix G)
- structured parameters of requested data (or samples) (based on the data model of the Directory [7]),
- unstructured parameters of requested data (or samples) and extent of the request.

Both accepted and rejected requests be logged by BBMRI-ERIC for minimum of 3 years. Data releases will be documented by BBMRI-ERIC and retained for minimum of 10 years.

If a researcher needs additional data that is not available in the centrally collected data set, or needs biological material, they can get in direct contact with the biobank. This access mode has no impact on data protection from CRC-Cohort perspective as BBMRI-ERIC is not involved. The BBMRI-ERIC Negotiator can be used to start negotiation directly with the contributing partner biobanks and access in such case is governed by normal BBMRI-ERIC Access Policy (Appendix H).

Accountability. For both Mode-2 and Mode-3, each access to the dataset is subject to logging of access. Minimum duration of log for 3 years for anonymized data (Mode-2) and 10 years for access to pseudonymized data (Mode-3).

3.1. Rationale of access modes

This section discusses rationale behind the access modes. The overview of the user interaction with the access modes can be found also in the Figure 2 on page 19.

Public/non-authenticated access (Mode-1). We start with a user (possible data requester) who has some interest in BBMRI-ERIC and its partner biobanks. The user may not have any medically relevant question, but may be, e.g., exploring extent of the BBMRI-ERIC infrastructure. Such a user is only limited to using publicly available services—in our particular case BBMRI-ERIC Directory. Here she can discover existence of the CRC-Cohort with very basic description of its parameters, such as juridical person, type and size of the cohort, covered diseases, and overall information on donors such as age range and their biological sex. The list of contributing partner biobanks will be also made available.

Supporting formulation of project ideas (Mode-2). When a researcher has a rough research idea and wants to check feasibility of the research, improve the study design, or even seeks

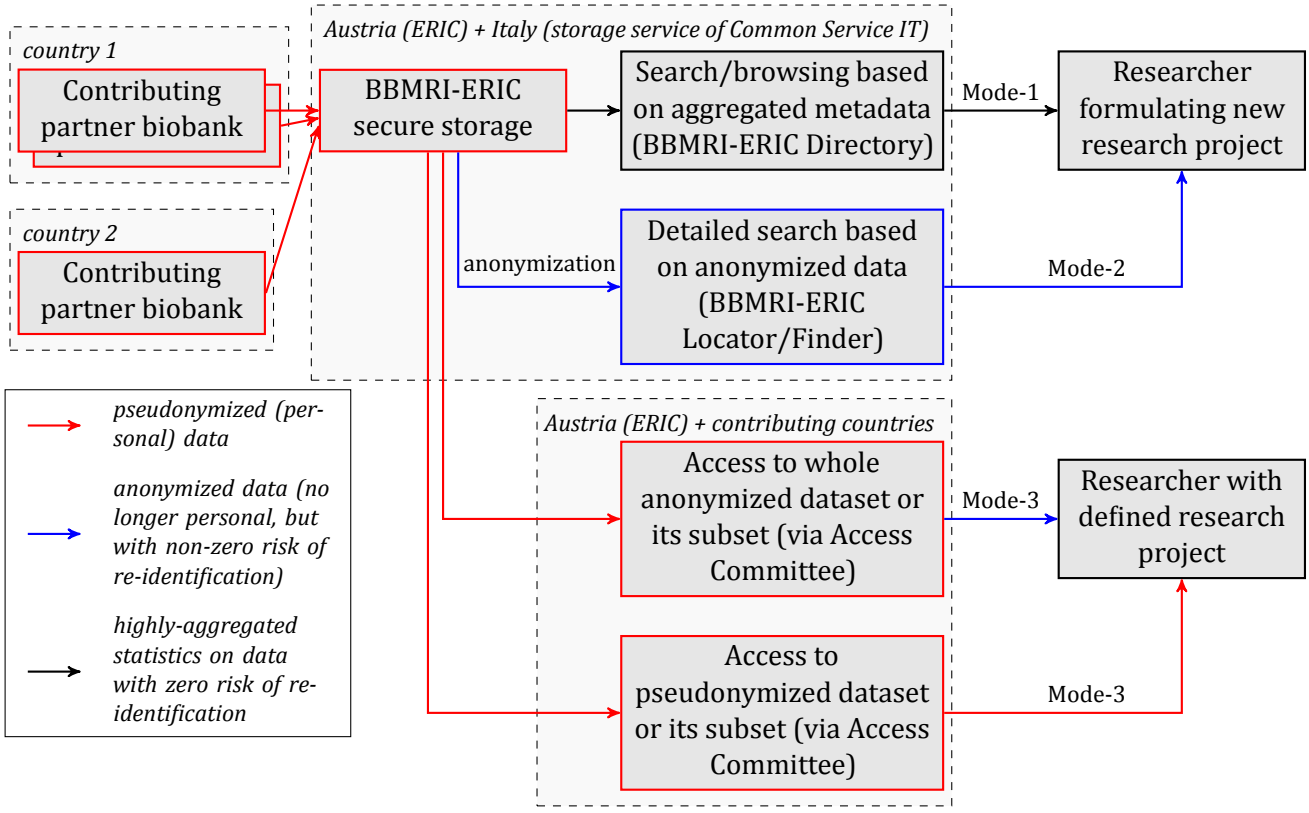


Figure 1: Overview of data processing within BBMRI-ERIC and provisioning of access. The access mode (e.g., Mode-1) refers to access modes defined in Section 3. Additional data and biological material may be obtained from the contributing partner biobanks.

to formulate a medically relevant project. To ensure implementability of the research, she needs to have estimates of available data and/or biological material, in order to estimate statistical power of the results. For this reason, BBMRI-ERIC will provide access to the anonymized data on estimates of data/sample availability using the BBMRI-ERIC Locator/Finder tool. The user can query different combinations of interested parameters and as the result, she receives a number of cases (if below threshold of k-anonymity with $k \geq 5$ considering all the data elements as pseudo-identifiers, suppression is implemented, i.e., zero results is reported).

Access to centrally collected pseudonymized/anonymized data set (Mode-3). A researcher, who already has well-defined and approved medically relevant project, can use the BBMRI-ERIC Negotiator tool to request access to the CRC-Cohort data set (Mode-3). Medically relevant research projects are typically granted access to the pseudonymized data set, as this represents the real data measures on individuals. However, as a part of the data minimiza-

tion principle,[18] the Access Committee can limit access also the anonymized data set if it is sufficient to reach the research goals. We envision this to be useful, e.g., for testing machine-learning approaches to support diagnoses processes. However, any positive results should be later validated on the pseudonymized data set before really considered meaningful.

As a part of requesting access via Mode-3, BBMRI-ERIC needs to ensure that the project is ethically acceptable. For this purpose the requester is offered an option to provide results of the ethics vote on the project. This serves as input for BBMRI-ERIC Ethics Check procedure (Appendix G); if the ethics vote is found sufficient, the BBMRI-ERIC Ethics Check procedure runs in the expedited mode to avoid duplicate ethics reviews. If the ethics check is not provided or if it is not sufficient, the BBMRI-ERIC offers ad hoc ethics review, see Appendix G. This can be understood also as a service for the requesters who do not have access to acceptable ethics review committee (e.g., when their home institution does not have one).

This access mode allows for efficient release of the data. Researcher signs only a single DTA with BBMRI-ERIC (instead of signing a custom DTA with each contributing partner biobank). Furthermore, BBMRI-ERIC guarantees a time limit on delivering decision of the Access Committee on access request.

The difference between Mode-2 and Mode-3 is that Mode-2 only allows querying the anonymized data set for authenticated users (no data transfer and no access to pseudonymized data), after accepting Terms and Conditions of Use of BBMRI-ERIC Services (Appendix F) preventing the abuse of the service and re-identification and privacy intrusion attempts. Using Mode-3, on the other hand, the user may receive the whole requested pseudonymized or anonymized data set, after positive decision of Access Committee and after the user accepts DTA (Appendix E).

[18] Rec.39; Art.5(1)(c) of GDPR [9].

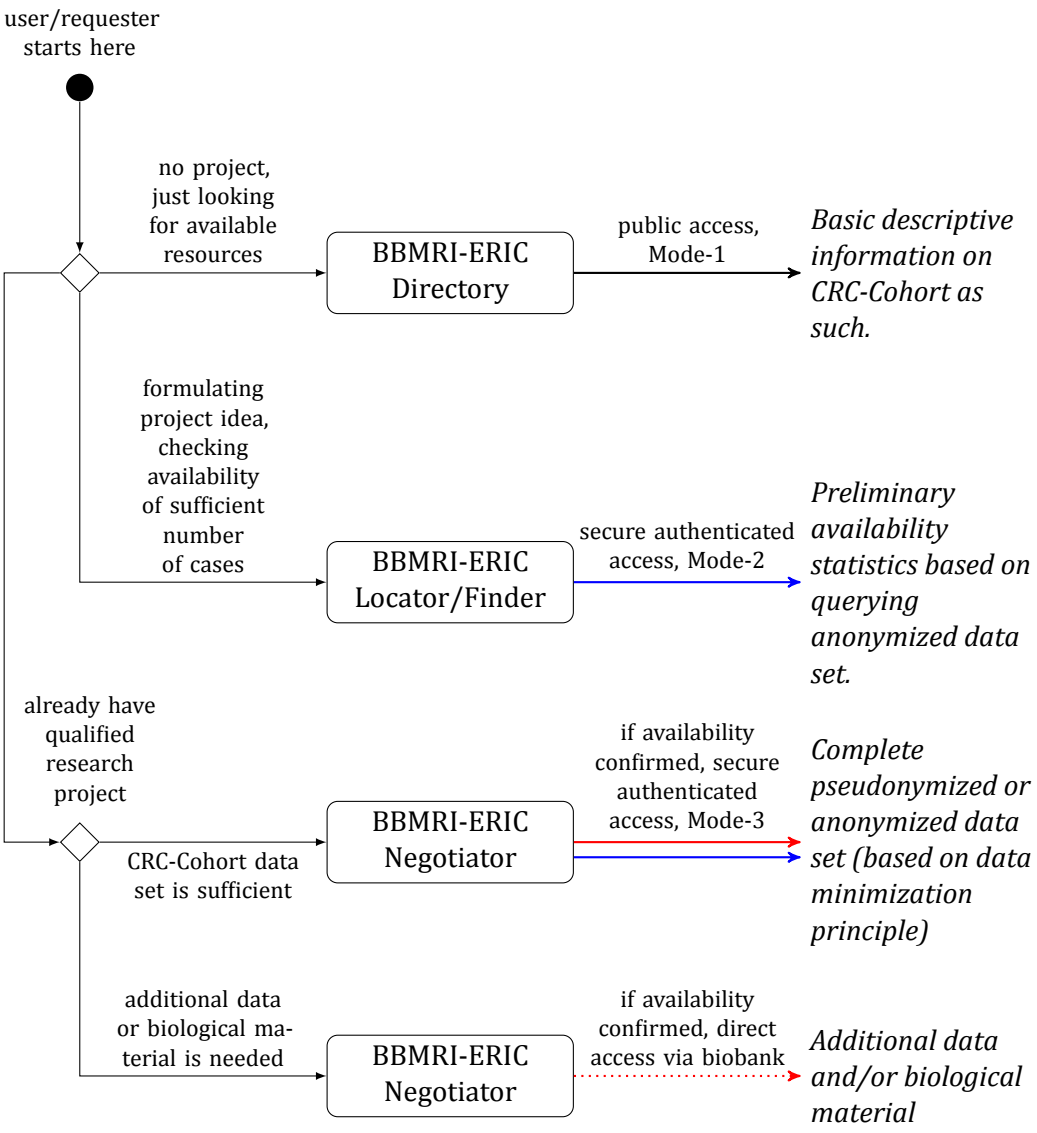


Figure 2: Schematics of user access mode rationale.

4. Implementation of Data Collection and Data Access

As shown in the Figure 3 on page 24, the data collection and access to data is implemented using several IT services provided by BBMRI-ERIC:

BBMRI-ERIC Authentication and Authorisation Infrastructure (AAI) has been developed to implement authentication and authorization consistently across other BBMRI-ERIC IT services, in compliance with BBMRI-ERIC security and privacy requirements [1] and architecture [12]. It utilizes eduGAIN federated authentication infrastructure[19] provided by GÉANT. A physical person can link their identities from. Authorization relies on BBMRI-ERIC's own instance of Perun [13], state-of-the-art authorization system used across biomedical research infrastructures.

Availability: Since April 2017. Version 1.0.

URL: `https://perun.bbmri-eric.eu/`

Authentication & authorization: Authentication via AAI itself. Initial authorization is decided by Senior IT & Data Protection Manager of BBMRI-ERIC, representatives of biobanks/collections are delegated capability to manage their own biobank.

CCDC Dedicated IT service operated by BBMRI-ERIC to collect the data for the CRC-Cohort. It features web-based interface for manual entry of the data and import REST-based API for ingesting data process by (semi-)automated extraction from the contributing partner biobanks. The output of the CCDC is a integrated pseudonymized data set comprising data from all the contributing partner biobanks stored on a secure storage. The security measure of data transfer and storage are discussed in Section 2.

Availability: Since 2016. Version 1.0.

URL: `https://ccdc.bbmri-eric.eu/`

Authentication & authorization: Authentication via BBMRI-ERIC AAI, authorization granted to the representatives of the collections from the contributing partner biobanks.

BBMRI-ERIC Directory Public IT service provided by BBMRI-ERIC with highly aggregated data describing biobanks, their sample and/or data collection, biobank networks, and relevant contacts. Details of data structures are provided in [7, 8] and Appendix C.

Note that BBMRI-ERIC Directory does not allow for querying properties of individual samples or donors; querying combination of properties for collection does not guarantee that appropriate combination is available on donor/sample level (e.g., searching for C18 diagnosis and FFPE material type ensures that collection has donors with C18 diagnosis and FFPE samples, but does not ensure that it has FFPE material of C18 diagnosis). This limitation is inherent to querying highly-aggregated data set

[19] `https://www.edugain.org/`

only and servers as pre-filtering of candidate biobanks that might possibly have relevant data or material, in order to receive more detailed availability information via BBMRI-ERIC Negotiator.

Availability: Since 2015. Version 1.0 in 2015, version 2.0 in 2015, version 3.0 in 2016, version 3.1 and 3.2 in January/July 2017, version 4.0 scheduled for Fall 2017.

URL: `https://directory.bbmri-eric.eu/`

Authentication & authorization: No authentication/authorization.

Query: Any combination of parameters of the data model in Appendix C. Typical query (demonstrating inclusion criteria for the CRC-Cohort):
`(diagnosisAvailable==C18.1 OR diagnosisAvailable==C18.2 OR diagnosisAvailable ==C18.3 OR diagnosisAvailable==C18.4 OR diagnosisAvailable=C18.5 OR diagnosisAvailable=C18.6 OR diagnosisAvailable=C18.7 OR diagnosisAvailable ==C19 OR diagnosisAvailable==C20) AND materialStoredTissueFFPE==true`

Response to query: List of data and/or biological material collections fulfilling the query criteria (interlinked to their parent biobanks, contact information, and biobank networks in which they participate).

Integration of CRC-Cohort: The highly aggregated information about the CRC-Cohort will be introduced into the BBMRI-ERIC Directory as a collection directly under BBMRI-ERIC. A new biobank network will be created with all the contributing partner biobanks added to it, to show the clear link.

BBMRI-ERIC Locator/Finder & Connector Locator is an IT service under development by BBMRI-ERIC, designed to allow search for properties of samples and their donors using federated search mechanism. This means the queries are distributed to participating biobanks using Connector operated by each biobank; the Connector uses harmonized data model and prepares candidate responses to the queries, which can be approved or modifies or even discarded by the biobanks, before returning them to the central Locator. Hence the biobanks retain full control of the responses. Locator is only available to authenticate users who agree to Terms and Conditions of Use of BBMRI-ERIC Services (Appendix F). The user may or may not have a project when using Locator.

From the data protection perspective, the default operation of the Connector is on pseudonymized data set. These partial results from individual Connectors of participating biobanks are collected into resulting data set, which is anonymized prior to being delivered to the user (requester). This "late anonymization" allows to apply less harm to the data (generalization, noise modulation, or suppression) to achieve the same resulting residual risk compared to Connector operating on only on anonymized data sets. However, it is possible to operate also on anonymized data sets and in case when only the CRC-Cohort is queried, either approach to anonymization yields equivalent results.

Availability: Scheduled 2018.

URL: (not yet available)

Authentication & authorization: Authentication by BBMRI-ERIC AAI, authorization by accepting Terms and Conditions of Use of BBMRI-ERIC Services (Appendix F).

Query: Any combination of sample data or donor data developed as a part of common Locator data model. While it is still under development, the most extensive scenario (from data protection perspective) would be ability to query any parameters in the CRC-Cohort data model (Appendix B).

Response to query: Total number and per-country number of found samples/donors.

Integration of CRC-Cohort: A dedicated Connector will be installed as a part of IT at BBMRI-ERIC, acting on the CRC-Cohort data set. Response approval will be done by the data manager of the CRC-Cohort.

BBMRI-ERIC Negotiator IT service provided by BBMRI-ERIC to obtain detailed sample/data availability information from the biobanks. It utilizes structured data query from the BBMRI-ERIC Directory (and later from the BBMRI-ERIC Locator/Finder once operational) to identify candidate biobanks that might possibly have relevant samples. In the Negotiator, the requester can add additional (non-structured) requirements on data/samples. Each request in the Negotiator requires already available relevant research project, so that biobanks can assess availability of informed consent on the material/data or permissions for secondary use of it. As a part of the BBMRI-ERIC Negotiator, the

Availability: Since May 2017. Version 1.0.

URL: `https://negotiator.bbmri-eric.eu/`

Authentication & authorization: Authentication by BBMRI-ERIC AAI. As condition necessary but not sufficient per se, authorization requires accepting Terms and Conditions of Use of BBMRI-ERIC Services (Appendix F). Authorization to pose queries requires providing information on the research project. Authorization to act on behalf of the collections require validation procedure between BBMRI-ERIC and the biobank.

Example of query: Each request comprises of:

- structured data, e.g.,
  `(diagnosisAvailable==C18.1 OR diagnosisAvailable==C18.2 OR diagnosisAvailable==C18.3 OR diagnosisAvailable==C18.4 OR diagnosisAvailable=C18.5 OR diagnosisAvailable=C18.6 OR diagnosisAvailable=C18.7 OR diagnosisAvailable==C19 OR diagnosisAvailable==C20)`
- unstructured data, e.g.,
  "The data need semantic annotation of the data in RDF format in order to enable biomarker discovery based on deep learning techniques (machine learning based on neural networks). Indicated data set size leads to rough estimate of 100 CPU/years capacity needed."
- project description, including results of the previous ethics vote if available.

Example of response to query: "There is 10,000 cases available in the CRC-Cohort data set. The approval has been obtained from all the contributing partner biobanks, which form the Access Committee. The data access requires signing DTA between you and BBMRI-ERIC, as shown in Appendix E. After signing the DTA, you will be granted access to downloading encrypted data after authorization via BBMRI-ERIC AAI." (see enabling download of the data over HTTPS below)

Integration of CRC-Cohort: Making the CRC-Cohort findable in the BBMRI-ERIC Directory and upcoming BBMRI-ERIC Locator/Finder directly ensures it becomes available also in the Negotiator. The data manager of CRC-Cohort will be authorized representative to act on behalf of the CRC-Cohort.

HTTPS data download with TLS 1.1 or newer Implementation of access to the pseudonymized or anonymized data set for download (Mode-3) under DTA using secure authenticated access. The data access will be available for authenticated and authorized users only, over secure communication channel (TLS 1.1 or newer). As a second factor of authenticity, the provided data will be encrypted: either using asymmetric encryption to allow the requester to use her private key of X.509 certificate, or symmetrically using one-time key generated for the particular release of data delivered to the user by an independent secure channel. Release of the data will use a unique release ID and made accessible for the user under this ID.

Availability: Scheduled 2018.

URL: (not yet available)

Authentication & authorization: Authentication via BBMRI-ERIC AAI. Specific authorization for a user and data release using unique release ID.

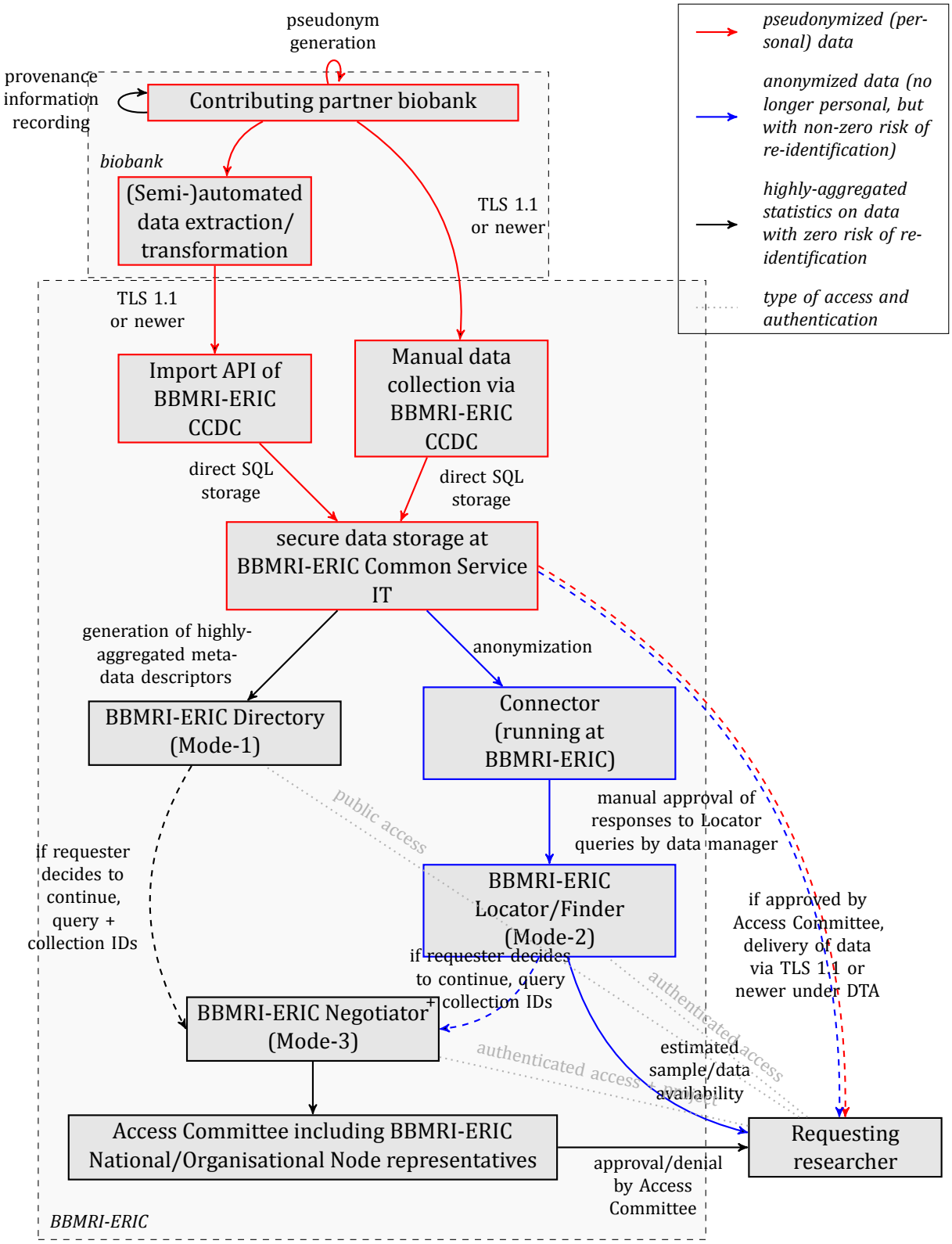


Figure 3: Detailed component-based view of data processing.

5. Overview of Data Protection/Security Measures

Overview of the whole data collection, storage, and access (including findability) is summarized in Figure 1. This diagram also shows types of data involved in the each step relevant from the data protection perspective. Figure 3 provides more detailed view of components and data transmissions involved.

Summary of data protection measures:

- *Storage of pseudonymized data* is secured using technical and organizational measures (including restricted access to physical storage infrastructure and encryption of storage, see Section 2.4 for more details) at BBMRI-ERIC as data controller as a part of BBMRI-ERIC Common Service IT.

- *Access to pseudonymized data* is controlled by Access Committee, see Section 3 Mode-3. This includes verification of requester's identity (authentication) and checking compliance of the project with informed consent or approval for secondary use of data. DTA (see Appendix E) requires for any data processor will require the same level of security as for storing the data by BBMRI-ERIC (see Section 2.4). The pseudonyms are generated anew for each release of the data to avoid straightforward attacks based on reuse of the pseudonym.

- Current approach prevents (privacy-preserving) record linkage. In the future, once EUPID[20] [14] is generally accepted across the participating European countries, the pseudonym generation policy may be updated.

- Data will be *anonymized* as a part of sample-level availability statistics provided via BBMRI-ERIC Locator/Finder tool, with minimum anonymization being k-anonymity with $k \geq 5$ considering all the data elements as pseudo-identifiers (see Section 3, including discussion of review of anonymization under GDPR). Anonymization can also occur in access Mode-3, if the data minimization principle[21] leads to using anonymized data instead of pseudonymized data. Anonymized data will not be stored as such permanently, the anonymization is applied as a part of query response (Mode-2) or as a part of data release (Mode-3).

  Because of non-zero re-identification risk, this data will be only available for authenticated users after accepting conditions of use excluding re-identification attempts (both Mode-2 and Mode-3).

  The utilized minimum anonymity is not necessary resilient to certain type of attacks (particularly inference attacks in case of k-anonymity, see summary in [4]). This has been intentionally chosen in order to balance utility of the data vs. privacy protection. Introducing additional requirements, such as l-diversity, would result in further hiding data, or generalizing the data values, or introducing artificial noise into the data set, and thus substantially worsening utility of the data set. For the same reason, the

[20] `https://eupid.eu/`

[21] Rec.39; Art.5(1)(c) of GDPR [9].

anonymization will be done centrally by BBMRI-ERIC on the data sets to be released: doing anonymization on the larger data set after aggregating it from the contributing partner biobanks results in less damage to the data, while ensuring the same residual risk of re-identification. As a complementary organizational/technical measure, any access to anonymized data in the CRC-Cohort requires authentication and acceptance of Terms and Conditions of Use of BBMRI-ERIC Services (Appendix F), which forbid re-identification attempts. Hence the anonymized data set is not published as fully non-personal data and take corresponding risks into account.

Anonymization will be reviewed periodically to assess its adequacy and stronger anonymization techniques may be applied when found inadequate with respect to state-of-the-art and legal and ethical requirements.

- *Network transport* of pseudonymized and anonymized data will be done via secure network transport protocols only (TLS 1.1 or newer with X.509 server certificate is issued by commonly accepted root certificate authority).

6. Risk Analysis

The risk analysis in this section is based on STRIDE [15] and LINDDUN [16] approaches. The overall *risk level* is qualitatively assessed using *likelihood of a threat* and *level of impact* as shown Table 1.

Likelihood of a threat	Level of impact		
	Low (+)	Medium (++)	High (+++)
Low (+)	+	+	++
Medium (++)	+	++	+++
High (+++)	+	++	+++

Table 1: Qualitative risk assessment.

Table 2: Risk assessment for threats (STRIDE and LINDDUN) to data flows.

Data flow threats	*Example*	*Risk*				*Countermeasure*
		Collection	Storage	Access Mode-2	Access Mode-3	
Tampering	Malicious modification of data or code, e.g., by man-in-the middle attack possible because of weak message or channel integrity checks	+++	+++	+	+++	Secure data communication (TLS 1.1 or newer)
Information disclosure	Exposure of data to unauthorized persons, e.g., by man-in-the-middle because of lack of confidentiality for the channel	+++	+++	+	+++	
Denial of service	Consumption of large quantities of fundamental resources due to weak message or channel integrity	+	++	+	+	
– (not relevant), + (low), ++ (medium), +++ (high)						

Table 3: Risk assessment for security (STRIDE) threats to storage and processing.

Security threat	*Example*	*Risk*				*Countermeasure*
		Collection	Storage	Access Mode-2	Access Mode-3	
Spoofing	Pose as something or somebody else	+++	+++	+	+++	Authentication system, configuration management
Tampering	Malicious modification of data or code	+++	+++	+	+++	Authorization system
– (not relevant), + (low), ++ (medium), +++ (high)						

Continued on next page...

... continued from previous page.

Security threat	*Example*	*Risk*				*Countermeasure*
		Collection	Storage	Access Mode-2	Access Mode-3	
Repudiation	Denial of having received data	+++	+++	++	++	Auditing and logging
Information disclosure	Exposure of information to unauthorized individuals	+++	+++	+	+++	Authorization System, Input Validation
Denial of service	Resources are not available due to overload or attack	++	++	+	+	Configuration management, input validation
Elevation of privilege	A user gains unauthorized access to resources	+++	+++	+	+++	Authorization system
– (not relevant), + (low), ++ (medium), +++ (high)						

Table 4: Risk assessment for privacy (LINDDUN) threats to storage and processing.

Privacy threat	*Example*	*Risk*				*Countermeasure*
		Collection	Storage	Access Mode-2	Access Mode-3	
Linkability	Possibility to detect that different data items are related to the same entity	++	++	+	++	Pseudonymization (Sections 2.2 and 2.3 and anonymization (Section 3 Mode-2). Encryption, access control system.
Identifiability	Possibility to relate a set of data to a specific entity / person; to recognize a person by characteristics	++	++	+	++	
Content unawareness	A patient is unaware of the information used/shared by the system	+	++	+	++	Informed consent management
Policy/consent non-compliance	Lack of evidence that data shared by the system meets applicable legal, policy or consent requirements	+	++	+	++	Legal regulations, informed consent mgmt., data provider forms, ethics committee approval, data access comm. approval, DTA/MTA.
– (not relevant), + (low), ++ (medium), +++ (high)						

Glossary

AAI Authentication and Authorisation Infrastructure. 20, 22, 23

Access Committee Access Committee for CRC-Cohort, see Appendix I. 14, 15, 17, 18, 23–25, 85

ADOPT BBMRI-ERIC project Project run by BBMRI-ERIC supported by the Horizon 2020 grant number 676550 (implementAtion anD OPeration of the gateway for healTh into BBMRI-ERIC – ADOPT BBMRI-ERIC). 4–8, 14

API Application Programming Interface. 11, 20, 24, 29

Assembly of Members Assembly of representatives of the member countries of BBMRI-ERIC. 77

BBMRI-ERIC Directory Information service by BBMRI-ERIC, providing highly aggregated data about the biobanks and their collections of biological material and data. Previously also known as BBMRI Catalogue. `https://directory.bbmri-eric.eu/`. 2, 14–17, 19–24, 44

BBMRI-ERIC Locator/Finder Services for searching preliminary availability information on samples and data sets based on privacy-preserving federated querying. The services constitute BBMRI-ERIC Federated Platform.. 14, 15, 17, 19, 21–25, 72

BBMRI-ERIC National/Organisational Node National Nodes means an entity, not necessarily with legal capacity, designated by a Member State, that coordinates the national Biobanks and Biomolecular Resources, and links its activities with the pan- European activities of BBMRI-ERIC. Organisational Node means an entity, not necessarily with legal capacity, designated by an intergovernmental organisation that coordinates the Biobank(s) and Biomolecular Resources of the organisation, and links its activities with those of the pan-European infrastructure, BBMRI-ERIC. 4, 24, 30

BBMRI-ERIC Negotiator A service for facilitating access to BBMRI-ERIC partner biobanks, by orchestrating and simplifying the communications between researchers (requesters) and biobankers. `https://negotiator.bbmri-eric.eu/`. 15–17, 19, 21, 22, 24, 72

CCDC Colorectal Cancer Data Collection software tool developed by BBMRI-ERIC for manual collection of CRC-Cohort data, featuring also import API for ingesting data prepared using (semi-)automated processing. See `https://ccdc.bbmri-eric.eu/`. 10–12, 20, 24

Common Service A Common Service means a facility of BBMRI-ERIC according to Article 15(1) according to the Statutes.. 29

Common Service IT Common Service on Information Technologies (IT). 12, 13, 17, 24, 25

Common Service ELSI Common Service on Ethical, Legal, and Societal Issues. 85

CRC-Cohort Colorectal Cancer Cohort, the subject of this document. 1–7, 9–11, 13–17, 19–23, 26, 29, 36, 38, 85

DOI Digital Object Identifier, `https://www.doi.org/`. 13

DTA Data Transfer Agreement. 3, 9, 15, 18, 23–25, 34, 36, 57, 64, 85

FFPE Formaline-fixed paraffin-embedded (type of archival tissue). 10, 20

GDPR General Data Protection Regulation. 7, 8, 15, 18, 25
GÉANT The GÉANT Association, http://www.geant.net/. 11

IMI CoP IMI CODE of PRACTICE on SECONDARY USE of MEDICAL DATA in SCIENTIFIC RESEARCH PROJECTS[22]. 8, 34
ISO International Organization for Standardization (ISO), http://www.iso.org/. 30
IT Information technology. 6, 20–22

LINDDUN Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance. See [16]. 4, 27, 28

Member EU Member States, third countries as well as intergovernmental organisations may become Members BBMRI-ERIC. 29, 77
MIABIS Minimum Information About BIobank data Sharing (MIABIS), community standard for representing biobanks and their contents. See https://github.com/MIABIS/miabis/wiki. 44, 45
MTA Material Transfer Agreement. 36

partner biobank Biobanks partnering with BBMRI-ERIC via BBMRI-ERIC National/Organisational Node, as defined in the Statues of BBMRI-ERIC, Article 1, §10 [2, Annex 1]. 4, 6–11, 13, 15–18, 20, 21, 23, 24, 26, 34, 35, 85

REST Representational state transfer (REST) or RESTful web services, see https://en.wikipedia.org/wiki/Representational_state_transfer. 11, 20

SPR BBMRI-ERIC Security & Privacy Requirements [1]. 8
SQL Structured Query Language, [17]. 11
STRIDE Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege. See [15]. 4, 27

TLS Transport Level Security. 11, 23, 24, 26, 27

X.509 International Organization for Standardization (ISO) recommendation X.509, see http://www.itu.int/rec/T-REC-X.509/en. 11, 26

[22] https://www.etriks.org/wp-content/uploads/2015/11/Code-of-Practice-on-Secondary-Use-of-Medical-Data-with-recognition.pdf

References

[1] P. Holub et al. *Security and privacy requirements and secure storage architecture*. Delivered as Milestone M6.2 document of EGI- Engage. Mar. 2016. DOI: 10.5281/zenodo.161551. URL: https://doi.org/10.5281/zenodo.161551.

[2] European Commission. "COMMISSION IMPLEMENTING DECISION of 22 November 2013 on setting up the Biobanks and Biomolecular Resources Research Infrastructure Consortium (BBMRI-ERIC) as a European Research Infrastructure Consortium (2013/701/EU)". In: *Official Journal of the European Union* L 320/63.November (2013), pp. 63–80. URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:320:0063:0080:en:PDF.

[3] A. Cambon-Thomsen. *Ethical Governance Framework and Ethics Check for ADOPT BBMRI-–ERIC*. ADOPT BBMRI-ERIC Deliverable D5.7. Mar. 2016. DOI: 10.5281/zenodo.823029. URL: https://doi.org/10.5281/zenodo.823029.

[4] Article 29 Data Protection Working Party. *Opinion 05/2014 on Anonymisation Techniques*. Tech. rep. April. 2014, p. 37. URL: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216%7B%5C_%7Den.pdf.

[5] I. Fortier et al. "Maelstrom Research guidelines for rigorous retrospective data harmonization." In: *International journal of epidemiology* 46.1 (Feb. 2017), pp. 103–105. ISSN: 1464-3685. DOI: 10.1093/ije/dyw075. URL: http://ije.oxfordjournals.org/lookup/doi/10.1093/ije/dyw075%20https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyw075%20http://www.ncbi.nlm.nih.gov/pubmed/27272186%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5407152.

[6] T. Flanagan, I. Fortier, M. Fon Sing, and C. Moore. "An International Cross-cohort Harmonization and Data Integration Initiative towards Achieving Statistical Power and Meaningful Results". In: *International Journal for Population Data Science* 1.1 (Apr. 2017), p. 23889. ISSN: 2399-4908. DOI: 10.23889/ijpds.v1i1.384. URL: https://ijpds.org/article/view/384.

[7] P. Holub et al. "BBMRI-ERIC Directory: 515 Biobanks with Over 60 Million Biological Samples". In: *Biopreservation and Biobanking* 14.6 (2016), pp. 559–562.

[8] P. Holub et al. *Consolidated Registry of BBMRI-ERIC Biobanks*. Sept. 2016. DOI: 10.5281/zenodo.273550. URL: https://doi.org/10.5281/zenodo.273550.

[9] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. May 2016.

[10] *BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data*. Submitted for approval by BBMRI-ERIC Assembly of Members. 2017.

[11] G. Dagher et al. *Harmonised Access Procedure to Samples and Data*. ADOPT BBMRI-ERIC Deliverable D4.1. Sept. 2016. DOI: 10.5281/zenodo.823013. URL: https://doi.org/10.5281/zenodo.823013.

[12] P. Holub and Common Service IT. *Security and Privacy Architecture*. Sept. 2016. DOI: 10.5281/zenodo.159444. URL: https://doi.org/10.5281/zenodo.159444.

[13] M. Procházka, S. Licehammer, and L. Matyska. "Perun—Modern approach for user and service management". In: *IST–Africa Conference Proceedings, 2014*. IEEE. 2014, pp. 1–11.

[14] H. Ebner et al. “Piloting the European Unified Patient Identity Management (EUPID) Concept to Facilitate Secondary Use of Neuroblastoma Data from Clinical Trials and Biobanking”. In: *Stud Health Technol Inform* 223 (2016), pp. 31–38.

[15] M. Howard and S. Lipner. *The security development lifecycle: SDL-A process for developing demonstrably more secure software*. 2006.

[16] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen. “A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements”. In: *Requirements Engineering* 16.1 (2011), pp. 3–32.

[17] C. J. Date. *A guide to the SQL standard: a user’s guide to the standard database language SQL*. 4th ed. Addison-Wesley Professional, 1997. ISBN: 978-0-201-96426-4.

[18] R. Merino-Martinez et al. “Towards Global Biobank Integration by Implementation of the Minimum Information About Biobank data Sharing (MIABIS 2.0 Core)”. In: *Biopreservation and Biobanking* 14 (4 Aug. 2016).

[19] L. Norlin et al. “A minimum data set for sharing biobank samples, information, and data: MIABIS”. In: *Biopreservation and biobanking* 10.4 (2012), pp. 343–348.

[20] P. Puchois. “General overview of industry needs and requirements”. In: *HandsOn: Biobanks*. July 2015.

[21] H.-E. Wichmann et al. “Comprehensive catalog of European biobanks”. In: *Nature Biotechnology* 29.9 (2011), pp. 795–797.

A. Revisions of This Document

- **Version: 1.5**
  - Date: 2023-05-12
  - Author(s): Petr Holub
  - Reviewer(s):
  - Summary of changes: Added Finder next to the Locator, as the BBMRI-ERIC Federated Platform comprises both tools with the same purpose.
- **Version: 1.4**
  - Date: 2023-01-03
  - Author(s): Petr Holub
  - Reviewer(s):
  - Summary of changes: Added assigned DOI.
- **Version: 1.3**
  - Date: 2023-01-03
  - Author(s): Petr Holub
  - Reviewer(s):
  - Summary of changes: Fixed mistake in description of Appendix E (Data Transfer Agreement (DTA) Template (BBMRI-ERIC – Researcher))
- **Version: 1.2**
  - Date: 2020-03-20
  - Author(s): Ayodeji Adeniran
  - Reviewer(s): Petr Holub
  - Summary of changes: Updates with respect to GDPR terminology.
- **Version: 1.1**
  - Date:
  - Author(s): Petr Holub
  - Reviewer(s):
  - Summary of changes: Changed role of BBMRI-ERIC from "data processor" to "data controller" (reverting previously introduced change created major issues in contracts with contributing biobanks). Updated contract templates in Appendices D and E – added ADOPT BBMRI-ERIC logo. Updated data model based on comments and questions received from the biobanks.
- **Version: 1.0**
  - Date: 2017-10-30
  - Author(s): Petr Holub
  - Reviewer(s):

- Summary of changes: Changed role of BBMRI-ERIC from "data controller" to "data processor" (only making the role of BBMRI-ERIC more restricted, no changes in terms of procedures). Fixed mistake in inclusion criteria (diagnosis 18.0 is also eligible). Explicit marks of alignment of the Policy to the IMI CoP. Simplified image (page 7) of contracting model: Data Provider Agreement and Reimbursement Contract between partner biobank and BBMRI-ERIC are now one process (arrow). Explanation of the reimbursement model (normal delivery, expedite delivery, manual-only delivery for small biobanks – page 7).

- **Version: 0.9**
  - Date: 2017-09-13
  - Author(s): Petr Holub, Irene Schlünder, Outi Törnwall
  - Reviewer(s):
  - Summary of changes: Revised Data Provider Agreement template and Data Transfer Agreement template. Updated Director General of BBMRI-ERIC. Update of of EU funding recognition to comply with Grant Agreement art. 29.4. Explicit statement that it is reponsibility of the partner biobank to ensure that the data collection and transfer complies with relevant legal and ethical requirements. Documented transfer procedure to outside of EU. Added incidental findings handling and return of results. Added information that collected genetic information is not sufficient to single out individuals. Added logging of requests and data releases by BBMRI-ERIC. Minor document layout improvements.

- **Version: 0.8**
  - Date: 2017-07-16
  - Author(s): Michael Hummel, Kurt Zatloukal
  - Reviewer(s):
  - Summary of changes: Improvement of purposes of the cohort. Minor improvements of the text.

- **Version: 0.7**
  - Date: 2017-07-13
  - Author(s): Petr Holub
  - Reviewer(s):
  - Summary of changes: Summary of document structure, fixed version numbering. Updated link to BioMedBridges for templates of DTAs (Appendices D and E). Updated templates themselves to remove BioMedBridges logo. Added definitions into Terms & Conditions in Appendix F. Improvements of Figure 3.

- **Version: 0.6**
  - Date: 2017-07-12
  - Author(s): Irene Schlünder, Outi Törnwall, Petr Holub
  - Reviewer(s):

    - Summary of changes: Specification of access committee setup and operation. Overall improvements, readability improvements and simplifications.

- **Version: 0.5**
    - Date: 2017-07-06
    - Author(s): Petr Holub, Irene Schlünder
    - Reviewer(s):
    - Summary of changes: Clarifications required from legal perspective.

- **Version: 0.4**
    - Date: 2017-06-29
    - Author(s):
    - Reviewer(s): Marialuisa Lavitrano
    - Summary of changes: Minor language fixes, questions on ethics committe and acknowledgements in the DTA.

- **Version: 0.3**
    - Date: 2017-06-28
    - Author(s): Petr Holub
    - Reviewer(s):
    - Summary of changes: Better description of anonymization and related procedures. Improved GDPR compliance description. Publishing list of contributing partner biobanks to support informed consent withdrawal.

- **Version: 0.2**
    - Date: 2017-06-26
    - Author(s): Petr Holub, Kurt Zatloukal, Michael Hummel
    - Reviewer(s): Outi Törnwall
    - Summary of changes: More detailed description of purposes, clarification of text, high-level overview of access into a diagram.

- **Version: 0.1**
    - Date: 2017-06-19
    - Author(s): Petr Holub
    - Reviewer(s): –
    - Summary of changes: Initial version.

B. Data Model for CRC-Cohort

Colorectal cancer experts have defined a comprehensive data set, summarized in this Appendix, for the colorectal cancer cohort that would provide a sufficient pool of information for the researchers to query and explore if the cohort is suitable for particular research questions.

Working Group

- *Medicine:* Marialuisa Lavitrano, Michael Hummel, Kurt Zatloukal, Dalibor Valík, Olli Carpén, Gerrit Meijer, Rudolf Nenutil, Barbara Parodi, Annemieke Hiemstra, Mariska Bierkens, Geraldine Vink, Heiden Esmeralda
- *IT:* Petr Holub, Frank Ückert, Diogo Alexandre, Ondřej Vojtíšek, Rumyana Proynova

B.1. Inclusion criteria

The following consensus has been reached on the inclusion criteria (not directly part of the data model, but also necessary for correct interpretation of the resulting data set):

- Colorectal cancer as a primary diagnosis (C18.0 to C18.7, C19, C20)
- Available FFPE – surgical material
- Availability of all REQUIRED data
- 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. This assumes signing Material Transfer Agreement (MTA)/DTA.

Note: Biopsies do not qualify as surgical tumor material. Biopsies do not provide sufficient amount of material to support multiple research projects.

B.2. Overview of the data items

Defined variables:

- Sex
- Participation in clinical study
- Age at primary diagnosis
- Known risk factors for CRC
- Time of recurrence (metastasis) and availability of biological material from recurrence (if recurrence occurred)
- Vital status and survival information
  - Timestamp of last update of vital status
  - Overall survival status

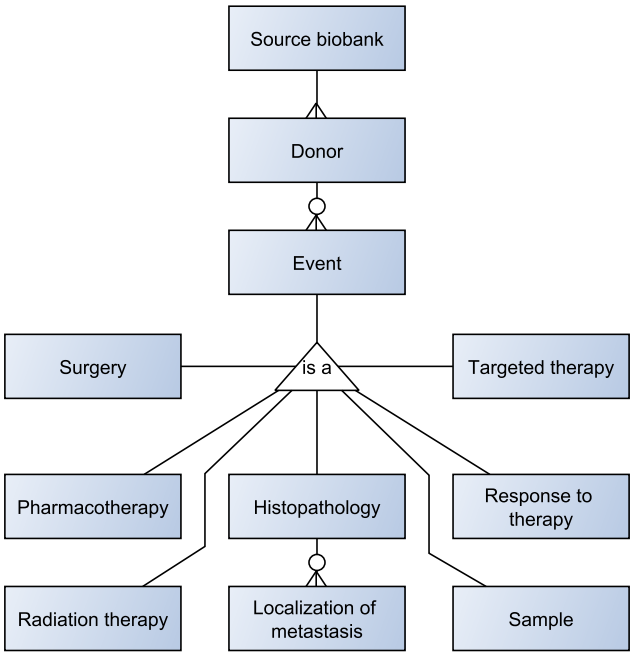
- Surgery
  - Time difference between initial diagnosis and surgery
  - Surgery radicality
  - Type of surgery
- Pharmacotherapy
  - REQUIRED if occurred
  - Date of start
  - Scheme of pharmacotherapy (including textual description of such scheme that does not fit into the provided list of schemes)
- Targeted therapy
  - REQUIRED if occurred
  - Date of start
  - Date of end
- Radiation therapy
  - REQUIRED if occurred
  - Date of start
  - Date of end
- Response to therapy
  - The response is linked to the patient and specified by a timestamp. This is to avoid need to specify to which therapy the response is, since there might be combination of different therapies.
  - Specific response
- Molecular markers (if applicable)
  - Microsatellite instability (if applicable)
  - Mismatch repair gene expression – IHC array for different genes (common for 3) (if applicable)
  - Risk situation (only HNPCC)
  - RAS mutation status (if applicable)
  - BRAF, PIC3CA. HER2 mutation status (optional)
- Histopathology part
  - TNM
  - UICC staging
  - WHO grading
  - Histological type of the cancer
  - Localization
  - Metastasis
  - High resolution (40×, i.e., $< 0.125\mu$m/pixel) digital image preferably from invasion front

    - Year of sample acquisition (indicative absolute date – ± 1 year accuracy is sufficient – to assess how old the material is)

- Diagnostic exam
    - Colonoscopy
    - Array of diagnostic methods (liver imaging, lung imaging, MRI, CT) done withing the context of primary diagnosis
    - CRC-Cohort does not collect data from these diagnostic methods, only information on their availability (i.e., they may be requested from the biobank by researchers)

All dates collected in the data model are relative (since initial diagnosis) to enhance privacy protection of patients, except for year of acquisition of samples (indicative year to assess how old samples are – ± 1 year accuracy is sufficient) and timestamp of last vital check (to assess freshness of data).

B.3. Entity-relation diagram



B.4. Tabular view of data model

(See next pages.)

Label	Level	Name	Data type [units] (validation)	Description	Ontology
Diagnostic exam					
DIAG_COLONOSCOPY	REQUIRED	Colonoscopy	LIST_OF_VALUES [Negative; Positive; Not done; Unknown]	Colonoscopy - Diagnostic exam. In case of rectal cancer, use rectoscopy also qualifies to answer TRUE here. But only rectoscopy in case of colon cancer does NOT qualify for TRUE. If the colonoscopy has been done outside of the biobank or the result is not available for some reason, the answer can be “not done”. This value shall be TRUE only if they were done within the context of the primary diagnosis. The values are advertising what is available in the biobank after further request and data is not provided as a part of collecting the central data set.	
DIAG_CT_DONE	REQUIRED	CT	LIST_OF_VALUES [Done, data available; Done, data not available; Not done; Unknown]	Diagnostic exam CT. This value shall be TRUE only if they were done within the context of the primary diagnosis. The values are advertising what is available in the biobank after further request and data is not provided as a part of collecting the central data set.	
DIAG_LIVER_IMAGING_DONE	REQUIRED	Liver imaging	LIST_OF_VALUES [Done, data available; Done, data not available; Not done; Unknown]	Liver imaging diagnostic exam. This value shall be TRUE only if they were done within the context of the primary diagnosis. The values are advertising what is available in the biobank after further request and data is not provided as a part of collecting the central data set.	
DIAG_X_DONE	REQUIRED	Lung imaging	LIST_OF_VALUES [Done, data available; Done, data not available; Not done; Unknown]	Lung imaging diagnostic exam. If CT or MRI or PET scan is available, this should be also considered one of the “Done” options. This value shall be TRUE only if they were done within the context of the primary diagnosis. The values are advertising what is available in the biobank after further request and data is not provided as a part of collecting the central data set.	
DIAG_MRI_DONE	REQUIRED	MRI	LIST_OF_VALUES [Done, data available; Done, data not available; Not done; Unknown]	MRI diagnostic exam. This value shall be TRUE only if they were done within the context of the primary diagnosis. The values are advertising what is available in the biobank after further request and data is not provided as a part of collecting the central data set.	
Histopathology					
		TNM	N/A	TNM	
		UICC staging	N/A	UICC staging	
		WHO classification	N/A	WHO classification	
DIGITAL_IMAGING_AVAILABILITY	OPTIONAL	Availability digital imaging	LIST_OF_VALUES [Can be generated; No; Readily available]	Do you have high-resolution digital imaging (corresponding to magnification 40x) from the histopatology?. Only scans of the surgical material should be considered here. The rationale is that smaller sections of the material (e.g., biopsies) do not contain sufficiently representative material for machine learning approaches. Resolutions should be <0.125um/pixel (this is more accurate description of 40x).	
DIGITAL_IMAGING_INVASION_FRO	OPTIONAL	Availability invasion front digital imagi	LIST_OF_VALUES [Can be generated; Invasion front not included; No; Readily available]	Do you have high-resolution digital imaging (corresponding to magnification 40x) containing invasion front from the histopatology?	
BIOLOGICAL_MATERIAL_FROM_RE	OPTIONAL	Biological material from recurrence av	YES_NO [] ((true|false|yes|no|f|t))	Biological material from recurrence available	
HIST_METASTASIS	REQUIRED	Localization of metastasis	LIST_OF_VALUES [Adrenals; Bone marrow; Brain; Hepatic; Lymph nodes; None; Osseous; Peritoneum; Pleura; Pulmonary; Skin; Others]	Histopathology part - Localization of metastasis. Multiple metastases can be added, each with its own location. This is intended for primary diagnosis only.	
HIST_LOCALIZATION	REQUIRED	Localization of primary tumor	LIST_OF_VALUES [C 18.0 - Caecum; C 18.1 - Appendix; C 18.2 - Ascending colon; C 18.3 - Hepatic flexure; C 18.4 - Transverse colon; C 18.5 - Splenic flexure; C 18.6 - Descending colon; C 18.7 - Sigmoid colon; C 19 - Rectosigmoid junction; C 20 - Rectum]	Histopathology part - Localization of primary tumor	
HIST_MORPHOLOGY	REQUIRED	Morphology	LIST_OF_VALUES [Adenocarcinoma; Adeonsquamous carcinoma; High-grade neuroendocrine carcinoma; Large cell neuroendocrine carcinoma; Medullary carcinoma; Micropapillary carcinoma; Mixed adenoneuroendocrine carcinoma; Mucinous carcinoma; Serrated adenocarcinoma; Signet-ring cell carcinoma; small cell neuroendocrine carcinoma; Spindle cell carcinoma; Squamous cell carcinoma; Undifferentiated carcinoma; Other]	Histopathology Part - Morphology. This is a mandatory part of histopathological diagnosis, therefore it should be available. If really not available, “Other” may be used, but it is a sign of insufficient data detail	
Histopathology - TNM					
TNM_DISTANT_METASTASIS	REQUIRED	Distant metastasis	LIST_OF_VALUES [M0; M1; M1a; M1b; M1c; MX]	TNM - Distant metastasis. It shall be interpreted as pTN - for tumor samples and biopsies, as the TN should come from the sample or biopsy. M may come from imaging (hence it may come from cTNM clinical assessment). Rationale: pTNM - is more reliable and should be available for tumors and biopsies	
TNM_PRIMARY_TUMOR	REQUIRED	Primary Tumor	LIST_OF_VALUES [T0; T1; T2; T3; T4; T4a; T4b; Tis; TX]	TNM Primary Tumor. It shall be interpreted as pTN - for tumor samples and biopsies, as the TN should come from the sample or biopsy. M may come from imaging (hence it may come from cTNM clinical assessment). Rationale: pTNM - is more reliable and should be available for tumors and biopsies	
TNM_REGIONAL_LYMPH_NODES	REQUIRED	Regional lymph nodes	LIST_OF_VALUES [N0; N1; N1a; N1b; N1c; N2; N2a; N2b; N3; NX]	TNM - Regional lymph nodes. It shall be interpreted as pTN - for tumor samples and biopsies, as the TN should come from the sample or biopsy. M may come from imaging (hence it may come from cTNM clinical assessment). Rationale: pTNM - is more reliable and should be available for tumors and biopsies	

Histopathology - UICC staging					
UICC_STAGE	REQUIRED	Stage	LIST_OF_VALUES [0; I; II; II A; IIB; IIC; III; IIIA; IIIB; IIIC; IV; IVA; IVB; IVC]	UICC Stage. The stages list is based on 8th edition, and backwards compatible with earlier editions.	
UICC_VERSION	REQUIRED	UICC version	LIST_OF_VALUES [4th. ed (used before 1998); 5th. ed (used 1998-2002); 6th. ed (used 2003-2009); 7th ed. (used 2010-2017); 8th ed. (used since 2017); Not known]	The version of the UICC system under which the staging was done	
Histopathology - WHO classification					
WHO_GRADE	REQUIRED	Grade	LIST_OF_VALUES [G1; G2; G3; G4; GX]	Grade. For Sweden “medium high” shall map to G3, and “low medium” shall map to G2. This has to be documented in the provenance information	
WHO_GRADE_VERSION	REQUIRED	WHO version	LIST_OF_VALUES [1st ed. (1979-1990); 2nd ed. (1991-2000); 3rd ed. (2001-2010); 4th ed. (used since 2011); Edition not known]	The version of the WHO classification system used	
Molecular markers					
		KRAS mutation status	N/A	KRAS mutation status	
BRAF_PIC3CA_HER_MUTATION_S	OPTIONAL	BRAF, PIC3CA, HER2 mutation statu	LIST_OF_VALUES [Mutated; Not mutated; Partial information available; Not done]	BRAF, PIC3CA, HER2 mutation status. If only 1 or 2 of the three mutation analyses have been done, the “Partial information available” value shall be selected	
MM_MICROSAT_INSTABILITY	REQUIRED	Microsatellite instability	LIST_OF_VALUES [no; yes; not done]	Microsatellites analysed BAT26, D17S250, D5S346, BAT40, D2S123 and BAT25. Image cytometry does not qualify for comparability reasons	
MM_MISMATCH_REPAIR_GE	REQUIRED	Mismatch repair gene expression	LIST_OF_VALUES [expression; loss of expression; not done]	Mismatch repair gene expression – IHC array for different genes (common for 3). Expression of MLH1, MSH2, PMS2 and MSH6	
MM_RISK_SITUATION_HNPCC	OPTIONAL	Risk situation (only HNPCC)	YES_NO [] ((true|false|yes|no|f|t))	Risk situation (only HNPCC), Amsterdam criteria	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2933058 http://
Molecular markers - KRAS mutation status					
MM_KRAS_MUTATION_NRAS_EX4	REQUIRED	NRAS exon 4 (codons 117 or 146)	LIST_OF_VALUES [Mutated; Not mutated; Not done]	NRAS exon 4 (codons 117 or 146) mutation status	
MM_KRAS_MUTATION_NRAS_EX3	REQUIRED	NRAS exon 3 (codons 59 or 61)	LIST_OF_VALUES [Mutated; Not mutated; Not done]	NRAS exon 3 (codons 59 or 61) mutation status	
MM_KRAS_MUTATION_NRAS_EX2	REQUIRED	NRAS exon 2 (codons 12 or 13)	LIST_OF_VALUES [Mutated; Not mutated; Not done]	NRAS exon 2 (codons 12 or 13) mutation status	
MM_KRAS_MUTATION_KRAS_EX4	REQUIRED	KRAS exon 4 (codons 117 or 146) mu	LIST_OF_VALUES [Mutated; Not mutated; Not done]	KRAS exon 4 (codons 117 or 146)	
MM_KRAS_MUTATION_KRAS_EX3	REQUIRED	KRAS exon 3 (codons 59 or 61)	LIST_OF_VALUES [Mutated; Not mutated; Not done]	KRAS exon 3 (codons 59 or 61) mutation status	
MM_KRAS_MUTATION_KRAS_EX2	REQUIRED	KRAS exon 2 (codons 12 or 13)	LIST_OF_VALUES [Mutated; Not mutated; Not done]	KRAS exon 2 (codons 12 or 13) mutation status	
Patient Data					
PATIENT_ID	REQUIRED	Patient pseudonym	TEXT [] ()	A pseudonym for the patient. The pseudonym has to be generated in compliance with the CRC-Cohort Data Protection Policy requirements (Section 2.2).	
AGE_AT_PRIMARY_DIAGNOSIS	REQUIRED	Age at diagnosis (rounded to years)	NATURAL_NUMBER [a] (0<=x)	Age at initial histopathological diagnosis (biopsy or surgical specimen of the primary tumor) rounded to years.	http://purl.bioontology.org/ontology/SNOMEDCT/423493009
SEX	REQUIRED	Biological sex	LIST_OF_VALUES [female; male; other]	Biological sex of the person, defined by chromosomes.	http://purl.obolibrary.org/obo/PATO_0020000
DATE_DIAGNOSIS	OPTIONAL	Date of diagnosis	DATE [] (ISO_8601_WITH_DAYS)	Date at which colon cancer was diagnosed for the first time. Histopathological diagnosis by biopsy or surgery qualifies as primary diagnosis	
CLINICAL_STUDY_PARTICIPANT	OPTIONAL	Participation in clinical study	YES_NO [] ((true|false|yes|no|f|t))	Participation in clinical study	
TIME_OF_RECURRENCE_RELATIV	OPTIONAL	Time of recurrence (metastasis diagn	NATURAL_NUMBER [week] (0<=x)	Weeks between primary diagnosis and diagnosed recurrence. If only months is available, conversion is weeks := months * 4. Any re-occurrence of cancer, be it a local re-occurrence, a lymph node metastasis, or a distant metastasis	
Pharmacotherapy					
PHARMACOTHERAPY_END_RELAT	REQUIRED	Date of end of pharamcotherapy	NATURAL_NUMBER [week] (0<=x)	End of the drug intake in weeks since initial diagnosis.	
PHARMACOTHERAPY_START_REL	REQUIRED	Date of start of pharamacotherapy	NATURAL_NUMBER [week] (0<=x)	Start of the drug intake in weeks since initial diagnosis.	
PHARMACOTHERAPY_SCHEME	REQUIRED	Scheme of pharmacotherapy	LIST_OF_VALUES [5-FU 1000 mg/m2 i.v. continuous infusion, day 1-5, weeks 1 and 5; 5-FU 225 mg/m2 i.v. continuous infusion, 5 days per week; 5-FU 325-350 mg/m2 + LV 20 mg/m2i.v. bolus, day1-5, weeks 1 and 5; 5-FU 400 mg/m2 + 100 mg i.v. bolus, d 1,2, 11,12,21,22; Capecitabine 800-825 mg/m2 bid po, day 1-5, together with radiation or continuously untill end of radiation; Only preoperatively (no standard): 5-FU 250 mg/m2 i.v. continuous infusion on days 1-13 nad 22-35 and oxaliplatin 50mg/m2 i.v. day 1,8,22 and 29; UFT (300-340mg/m2/day) and LV (22.5-90 mg/day) po continuously, 5(-7) days per week, together with radiotherapy; Other]	Scheme of pharmacotherapy. If the theraphy was terminated or changed (e.g., dosage reduced), “Other” shall be selected. Additional textual information should be provided in such a case, see PHARMACOTHERAPY_SCHEME_DESCRIPTION	https://academic.oup.com/annonc/article/23/10/2479/195121
PHARMACOTHERAPY_SCHEME_D	if(PHARMAC	Other pharmacotherapy scheme	TEXT [] ()	Other pharmacotherapy scheme. When Other option is selected for pharmacotherapy scheme, the plain text description shall be provided. The plain text must include at least the chemical compounds used, the dosage and timing is optional	
Radiation therapy					
RADIATION_THERAPY_END_RELA	REQUIRED	Date of end of radiation therapy	NATURAL_NUMBER [week] (0<=x)	End of the radiation therapy in weeks since initial diagnosis.	
RADIATION_THERAPY_START_RE	REQUIRED	Date of start of radiation therapy	NATURAL_NUMBER [week] (0<=x)	Start of the radiation therapy in weeks since initial diagnosis. For combined therapies, they should be entered as separate therapies.	
Response to thearapy					
THERAPY_RESPONSE_TIMESTAM	REQUIRED	Date response was obtained in weeks	NATURAL_NUMBER [] (0<=x)	Date response was obtained in weeks since initial diagnosis	
THERAPY_RESPONSE	REQUIRED	Specific response	LIST_OF_VALUES [Complete response; Partial response; Progressive disease; Stable disease]	Response to therapy - Specific response	
Sample					

SAMPLE_MATERIAL_TYPE	REQUIRED	Material type	LIST_OF_VALUES [Healthy colon tissue; Tumor tissue; Other]	Type of specimen	
SAMPLE_PRESERVATION_MODE	REQUIRED	Preservation mode	LIST_OF_VALUES [Cryopreservation; FFPE; Other]	The preservation mode for the specimen	
SAMPLE_ID	REQUIRED	Sample ID	TEXT [] ()	An identifier, unique within the biobank	
YEAR_OF_SAMPLE_COLLECTION	REQUIRED	Year of sample collection	NATURAL_NUMBER [years] (0<=x)	Calender year in which the sample was collected.	
Surgery					
SURGERY_LOCATION	REQUIRED	Location of the tumor	LIST_OF_VALUES [C 18.0 - Cecum; C 18.1 - Appendix; C 18.2 - Ascending (right); C 18.3 - Hepatic flexure; C 18.4 - Transverse colon; C 18.5 - Splenic flexure; C 18.6 - Descending (left); C 18.7 - Sigmoid; C 19 - Rectosigmoid; C 19.9 - Rectosigmoid; C 20 - Rectum; C 20.9 - Rectum]	Location of the tumor	
SURGERY_TYPE_OTHER	OPTIONAL	Other surgery type	TEXT [] ()	Surgery type, if not present on the list	
SURGERY_RADICALITY	REQUIRED	Surgery radicality	LIST_OF_VALUES [R0; R1; R2; RX]	Whether the surgery removed the entire tumor.	
SURGERY_TYPE	REQUIRED	Surgery type	LIST_OF_VALUES [Abdomino-perineal resection; Anterior resection of rectum; Endo-rectal tumor resection; Left hemicolectomy; Low anteroir colon resection; Pan-procto colectomy; Right hemicolectomy; Sigmoid colectomy; Total colectomy; Transverse colectomy; Other]	Surgery type	
SURGERY_START_RELATIVE	REQUIRED	Time difference between initial diagnc	NATURAL_NUMBER [week] (0<=x)	Time difference between initial diagnosis and surgery. Weeks between initial diagnosis and date of surgery. Pre-operatively treated cases (neoadjuvant therapy) are welcome, but there needs to be surgery later on anyway, to have also sufficient amount of biological material.	
Targeted therapy					
TARGETED_THERAPY_END_RELA	OPTIONAL	Date of end of targeted therapy	NATURAL_NUMBER [] (0<=x)	Targeted therapy - Date of end (weeks since initial diagnosis)	
TARGETED_THERAPY_START_RE	REQUIRED	Date of start of targeted therapy	NATURAL_NUMBER [] (0<=x)	Targeted therapy - Date of start (weeks since initial diagnosis)	
Vital status and survival information					
OVERALL_SURVIVAL_STATUS	REQUIRED	Overall survival status	NATURAL_NUMBER [week] (0<=x)	Weeks after first colon cancer therapy started for the given person. If the data is collected at the source in months only, the conversion should be weeks := months*4	
VITAL_STATUS_TIMESTAMP	if(VITAL_ST	Timestamp of last update of vital statu	DATE [] (ISO_8601_WITH_DAYS)	Timestamp of last update of vital status	
VITAL_STATUS	REQUIRED	Vital status	LIST_OF_VALUES [death due to colon cancer; death due to other reasons; death for unknown reasons; person is still alive; unknown]	Vital status	

B.5. Changes to the inclusion criteria

- C18.0 diagnosis has been added to the inclusion criteria (previously omitted by mistake).
- It has been clarified that only biopsies do not qualify (however, the initial clearly stated requirement for availability of surgical material).

B.6. Changes in the data model

The data model has been updated from the initial version distributed in July 2017 based on the feedback and questions from the biobanks in the period of July to December 2017. The updates were intentionally minimum in order to help biobanks in updating their existing data harmonization procedures.

Data model description now also includes entity-relation diagram (Section B.3 on page 38) to clarify cardinalities of relations (e.g., that multiple metastases and therapies can be provided).

- Data model now includes patient's pseudonym explicitly for clarity reasons as `PATIENT_ID` (this attribute was always present but hidden from the data model).
- Data model now uses shorter value names (without duplicating attribute name).
- List of added REQUIRED attributes:
  - `YEAR_OF_SAMPLE_COLLECTION` – please note this attribute is indicative of year to distinguish relatively old samples, ± 1 year error is not a problem
- List of added OPTIONAL attributes:
  - `BIOLOGICAL_MATERIAL_FROM_RECURRENCE_AVAILABLE` (only if `TIME_OF_RECURRENCE_RELATIVE` is set)
  - `PHARMACOTHERAPY_SCHEME_DESCRIPTION`
- List of attributes where list of values has been updated:
  - `HIST_LOCALIZATION` – added complete list of relevant ICD-10 codes, including C18.0 (which was added to the inclusion criteria).
  - `UICC_STAGE`, `WHO_GRADE`, `TNM_DISTANT_METASTASIS` `TNM_REGIONAL_LYMPH_NODES` `TNM_PRIMARY_TUMOR` – added missing values based from latest revision of the standard (including values needed for backward compatibility with older revisions).
  - `UICC_VERSION` – added 8^{th} revision (published after initial data model design).
  - `BRAF_PIC3CA_HER_MUTATION_STATUS` – added partial information available option.

  - DIAG_COLONOSCOPY – added missing values (“unknown”, “done, data not available”).

- List of attributes with updated/clarifying comments:
  - DIAG_COLONOSCOPY, DIAG_CT_DONE, DIAG_LIVER_IMAGING_DONE, DIAG_X_DONE, DIAG_MRI_DONE – done within the context of primary diagnosis, for advertising availability of further data in the biobank (i.e., not collected as a part of the central collection).
  - TNM_DISTANT_METASTASIS TNM_REGIONAL_LYMPH_NODES TNM_PRIMARY_TUMOR – clarification that TN is required from pTNM, M may come from cTNM (based on imaging).
  - TIME_OF_RECURRENCE_RELATIVE – clarification what recurrence means
  - PHARMACOTHERAPY_SCHEME – clarification wen to use PHARMACOTHERAPY_SCHEME_DESCRIPTION
  - RADIATION_THERAPY_START_RELATIVE – clarification how to enter combined therapies
  - HIST_MORPHOLOGY – clarification
  - WHO_GRADE – note on mapping legacy values for Sweden
  - DIGITAL_IMAGING_AVAILABILITY – more precise definition of $\times 40$ resolution ($< 0.125 \mu$m/pixel)
  - DATE_DIAGNOSIS – clarification for cases where biopsy served as primary diagnosis
  - MM_MICROSAT_INSTABILITY – clarification of non-acceptability of image cytometry
  - PHARMACOTHERAPY_SCHEME – added reference to source paper for the value list
  - THERAPY_RESPONSE – added reference to source paper for the value list

- Other changes
  - DIAG_LIVER_IMAGING_DONE – changed name of the attribute for naming consistency

C. BBMRI-ERIC Directory Data Model

To enable samples and data to be searched in a comparable way, the first development step was designing an extensible data model, that covers all three key components of biobanks: (a) *biological material and associated physical storage facilities*, (b) *data and associated data storage facilities*, and (c) *expertise of the biobankers*.

The core of the data model for the Directory 2.0 relies on to MIABIS 2.0 [18], a standard data model for biobanking, which is evolution of the previously published MIABIS model [19]. As shown in Figure 4, this includes the following basic entities:

- **biobanks** are the institutional units hosting collections of samples and data, as well as providing expertise and other services to their users. This entity does not contain directly any attributes related to the samples or data, which are implemented via links to the collections that are available in the given biobank.
- **collections** are containers for sample sets and/or data sets, with support for recursive creation of sub-collections (of arbitrary finite depth); here properties of the samples and data can be described in aggregated form such as sample counts, diseases, material types, data types, gender, etc.;
- **networks of biobanks** (not defined in the MIABIS 2.0), which may include either whole biobanks or even individual collections inside the biobanks;
- auxiliary **contact information** contact information attached to biobanks, collections and networks needed to get access to samples or data (which is defined centrally to minimize redundancy in the information model).

The data model has been defined in a modular way such that auxiliary classes can be added to suit the needs of biobank (sub)communities, such as to describe clinical, population, research study based, non-human, and standalone collections. Particularly clinical collections are used to enforce existence of attributes describing available diagnoses (which is optional for other types), as it is among the most common search criteria [20]. Standalone collections are used in the countries with legal requirements on institutionalized biobanks, if there are some collections that do not meet these requirements (yet).

Collections. To enable deterministic counts for samples we followed recommendation of MIABIS 2.0 [18] that (sub)collections are strictly based on the concept of *set partitioning*: for any collection containing countable (discrete) elements (such as samples/aliquots, images), each element must be exactly in one collection (partition) on the given level of recursion, and there must be no empty collections. This allows for straightforward aggregation: content of each parent entity, be it a collection or a biobank, is a sum of child entities – collections, sub-collections, etc.

The two main groups of material in the collections are *physical material (samples)* and *data*. While samples without data rarely make any sense, the opposite situation with biobanks storing only data is common in many fields, such as imaging biobanks in radiology.

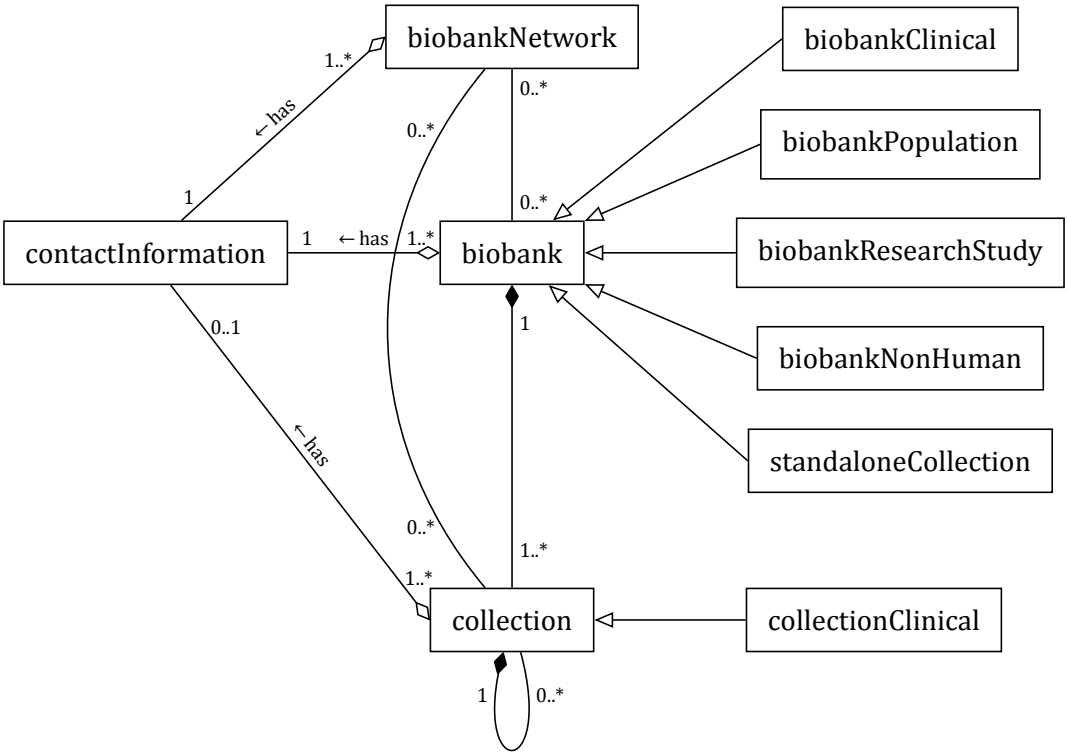


Figure 4: UML class diagram of the BBMRI-ERIC Directory Data Model.
Note that specialized biobank classes biobankClinical, biboankPopulation, biobankResearchStudy, and biobankNonHuman come as a legacy of collection-less Directory 1.0 and need to be reconsidered in the future development of the model, since these has become properties have been moved to the level of collections in MIABIS 2.0.

In practice, collections are most often created based on (a) purpose of collected material, or (b) life-cycle of the collected material, or (c) funding sources supporting the biobank.

Attributes of entities. Each of the entities have several sets of attributes, as detailed in Appendix C.1: (a) mandatory vs. optional parameters, (b) publicly visible parameters vs. parameters restricted for internal use by BBMRI-ERIC and its National Nodes.

Overall, the attributes can be summarized as follows:

- The attributes for biobanks focus on describing institutional aspects of biobanks, and are anticipated to be extended for other attributes such as available expertise and provided services. The attributes can be grouped into: *(a)* biobank ID (with possible support for mapping of various types of identifiers in the future), *(b)* type of the biobank [21], *(c)* contact information (via link to contactInformation object with as-

signed contactPriority, as well as URL), *(d)* head of the biobank and its institutional affiliation, *(e)* information about available information systems (restricted to internal purposes of BBMRI-ERIC and its National Nodes), *(f)* collaboration types supported.

- Attributes describing collections can be divided into:
  - organizational attributes: *(a)* collection ID (with possible support for mapping of various types of identifiers in the future), *(b)* contact information (via link to contactInformation object with assigned contactPriority, as well as URL), *(c)* head of the collection, *(d)* sample and data access policies.
  - attributes describing available physical material and its storage:*(a)* high-level view of stored material types (DNA, plasma, serum, urine, saliva, feces, RNA, blood, frozen tissue or equivalent, FFPE tissue or equivalent, immortalized cell lines), *(b)* type of collection (case control, cohort, cross sectional, longitudinal, twin study, quality control, population based, disease specific, birth cohort, other), *(c)* size of the collection (mandatory 10^n order of magnitude of collected discrete elements – typically samples,[23] with optional exact size with time stamp), *(d)* storage temperatures (based on SPREC 2.0 standard[24]).
  - attributes describing available data: *(a)* available data types (genealogical records, physiological/biochemical measurements, survey data, imaging, medical records), *(b)* access to other data sources (e.g., national registries).
  - attributes describing research participants: *(a)* sex and age of participants, *(b)* available diagnoses (with support for `?` and `*` wildcard characters replacing exactly one and zero or more characters respectively, to allow specification of whole classes of diagnoses, with appropriate search functionality in the Directory user interfaces).
- Biobank networks use attributes describing their institutional aspects as well as commonalities shared by the biobanks/collections participating in the given network.

  Participation of biobanks and collections in biobank networks is implemented via reference attributes from the biobanks and collections. Note *m:n* mapping between biobanks/collections and biobank networks, as one biobank/collection can participate in several biobank networks, and vice versa, each biobank network typically has more than one biobank/collection.

  The attributes can be grouped into: *(a)* biobank network ID (with possible support for mapping of various types of identifiers in the future), *(b)* contact information (via link to contactInformation object with assigned contactPriority, as well as URL), *(c)* commonalities of biobanks participating in the network (collection focus, charter, SOPs,

[23] We consider the search for exact number of samples meaningless before there is consensus on *sample and aliquot definition*, or having these terms standardized possibly as a part of ISO TC 276. We would also advise against abandoning these terms and using number of participants, as has already happened in some Nordic population biobanks, since such approach does not allow to differentiate between a biobank that collects one sample per participant and a time-consistent series of samples per each participant.

[24] `http://www.isber.org/?page=SPREC`

data and sample access policies, MTA/DTA, URL, or even complete representation where participating biobanks are only reachable via biobank network), *(d)* head of the biobank network and its institutional affiliation.

C.1. Detailed Description of Object Classes and Attributes

This section provides detailed overview of LDAP object classed defined in the BBMRI-ERIC Directory LDAP schema, together with description of their attributes. The tables with attribute description use the following shorthand notation:

Type Data type, where mapping to LDAP OID types is as follows:

OID	*Name*	*Note*
1.3.6.1.4.1.1466.115.121.1.15	string	case-insensitive substring search applied
1.3.6.1.4.1.1466.115.121.1.7	boolean	
1.3.6.1.4.1.1466.115.121.1.27	integer	
1.3.6.1.4.1.1466.115.121.1.50	phone	phone number
1.3.6.1.4.1.1466.115.121.1.11	country	two letter country code

C Cardinality, meaning how many times the attribute may be present. LDAP supports single-value and multi-value attributes (giving the upper limit on cardinality to 1 or n respectively), which may be further combined mandatory and optional status (giving the lower limit on cardinality to 1 or 0 respectively).

V Visibility, which can be *(i)* P … public, *(ii)* R … restricted to BBMRI-ERIC internal purposes.

C.2. contactInformation

Attribute	*Type*	*C*	*V*	*Description*
Mandatory				
contactID	string	1	P	Contact identifier.
contactEmail	string	1..n	P	Email according to MIABIS 2.0 – MIABIS-2.0-07-D.
contactCountry	country	1	P	Country according to MIABIS 2.0 – MIABIS-2.0-07-H.
Optional				
contactFirstName	string	0..1	P	First name according to MIABIS 2.0 – MIABIS-2.0-07-A.
contactLastName	string	0..1	P	Last name according to MIABIS 2.0 – MIABIS-2.0-07-B.

Continued on next page…

... continued from previous page.

Attribute	*Type*	*C*	*V*	*Description*
contactPhone	phone	0..n	P	Phone number according to MIABIS 2.0 including international prefix (+99999999999 form with no spaces) compliant also to E.123 norm – MIABIS-2.0-07-C.
contactAddress	string	0..n	P	Address according to MIABIS 2.0 – MIABIS-2.0-07-E.
contactZIP	string	0..1	P	ZIP according to MIABIS 2.0 – MIABIS-2.0-07-F.
contactCity	string	0..1	P	City according to MIABIS 2.0 – MIABIS-2.0-07-G.

C.3. collaborationStatus

Attribute	*Type*	*C*	*V*	*Description*
Optional				
collaborationPartners-Commercial	boolean	0..1	P	Biobank/collection can be used for collaboration with commercial partners.
collaborationPartners-Nonforprofit	boolean	0..1	P	Biobank/collection can be used for collaboration with non-for-profit partners.

C.4. biobank

Description of attributes also includes attributes of the superior objectClasses:

- collaborationStatus

Attribute	*Type*	*C*	*V*	*Description*
Mandatory				
contactIDRef	string	1..n	P	Reference to a contact ID.

Continued on next page...

... continued from previous page.

Attribute	*Type*	*C*	*V*	*Description*
contactPriority	integer	1	P	Priority of the contact 1..n (i.e., non-negative integer), where the highest priority should be used for contacting about given set of samples. E.g., if a collection has contactPriority=3, the biobank in which the collection resides has contactPriority=10, and the biobankNetwork to which the collection or biobank belongs has contactPriority=7, the biobank contact should be used.
biobankID	string	1	P	Unique biobank ID withing BBMRI-ERIC based on MIABIS 2.0 standard (ISO 3166-1 alpha-2 + underscore + biobank national ID or name), prefixed with bbmri-eric:ID: string – MIABIS-2.0-01.
biobankName	string	1	P	Biobank name according to MIABIS 2.0 – MIABIS-2.0-03.
biobankJuridicalPerson	string	1..n	P	Juristic person of a biobank according to MIABIS 2.0 – MIABIS-2.0-05.
biobankCountry	country	1..n	P	Country hosting the biobank according to MIABIS 2.0 – MIABIS-2.0-06.
biobankPartnerCharter-Signed	boolean	1	P	Biobank has signed BBMRI-ERIC Partner Charter.
Optional				
bioresourceReference	string	0..n	P	Bioresource reference to be cited when the bioresource (biobank/collection) is used for research.
biobankNetworkIDRef	string	0..n	P	Reference to a biobank network ID, to which the collection or biobank belongs; this attribute can also be used for biobank network, where it refers to the superior biobank network).
geoLatitude	string	0..1	P	Latitude of the biobank in the WGS84 system (the one used by GPS), positive is northern hemisphere.

Continued on next page...

... continued from previous page.

Attribute	*Type*	*C*	*V*	*Description*
geoLongitude	string	0..1	P	Longitude of the biobank in the WGS84 system (the one used by GPS), positive is to the East of Greenwich.
collaborationPartners-Commercial	boolean	0..1	P	Biobank/collection can be used for collaboration with commercial partners.
collaborationPartners-Nonforprofit	boolean	0..1	P	Biobank/collection can be used for collaboration with non-for-profit partners.
biobankITSupport-Available	boolean	0..1	R	Is IT support available at the biobank?
biobankITStaffSize	integer	0..1	R	Size of the biobank dedicated IT staff measured as 2^n.
biobankISAvailable	boolean	0..1	R	Has the biobank a computer-based Information System (IS)?
biobankHISAvailable	boolean	0..1	R	Has the biobank on-line or off-line connection to a Hospital Information System (HIS)?
biobankAcronym	string	0..n	P	Biobank acronym – MIABIS-2.0-02.
biobankDescription	string	0..n	P	Biobank description – MIABIS-2.0-08.
biobankURL	string	0..n	P	Biobank URL – MIABIS-2.0-04.
biobankHeadFirstName	string	0..n	P	First name of a person in charge of the biobank.
biobankHeadLastName	string	0..n	P	Last name of a person in charge of the biobank.
biobankHeadRole	string	0..n	P	Official role of the person in charge of the biobank: typically PI or Director.

C.5. collection

Description of attributes also includes attributes of the superior objectClasses:

- collaborationStatus

Attribute	*Type*	*C*	*V*	*Description*
Mandatory				

Continued on next page...

... continued from previous page.

Attribute	Type	C	V	Description
collectionID	string	1	P	Unique collection ID withing BBMRI-ERIC based on MIABIS 2.0 standard, constructed from biobankID prefix + :collection: + local collection ID string – MIABIS-2.0-01.
collectionName	string	1	P	Collection name according to MIABIS 2.0 – MIABIS-2.0-03.
materialStoredDNA	boolean	1	P	DNA: collection contains material of this type (MIABIS-2.0-14).
materialStoredPlasma	boolean	1	P	Plasma: collection contains material of this type (MIABIS-2.0-14).
materialStoredSerum	boolean	1	P	Serum: collection contains material of this type (MIABIS-2.0-14).
materialStoredUrine	boolean	1	P	Urine: collection contains material of this type (MIABIS-2.0-14).
materialStoredSaliva	boolean	1	P	Saliva: collection contains material of this type (MIABIS-2.0-14).
materialStoredFaeces	boolean	1	P	Faeces: collection contains material of this type (MIABIS-2.0-14).
materialStoredOther	string	1..n	P	Other: collection contains material of this type (MIABIS-2.0-14).
materialStoredRNA	boolean	1	P	RNA: collection contains material of this type (MIABIS-2.0-14).
materialStoredBlood	boolean	1	P	Blood: collection contains material of this type (MIABIS-2.0-14).
materialStoredTissue-Frozen	boolean	1	P	Frozen Tissue without formalin fixation or equivalent: collection contains material of this type (MIABIS-2.0-14).
materialStoredTissueFFPE	boolean	1	P	Tissue, formalin fixated and paraffin preserved or equivalent: collection contains material of this type (MIABIS-2.0-14).
materialStored-ImmortalizedCellLines	boolean	1	P	Immortalized cell lines: collection contains material of this type (MIABIS-2.0-14).

Continued on next page...

... continued from previous page.

Attribute	Type	C	V	Description
materialStoredIsolated-Pathogen	boolean	1	P	Isolated Pathogen: collection contains material of this type (MIABIS-2.0-14).
collectionTypeCase-Control	boolean	1	P	A case-control study design compares two groups of subjects: those with the disease or condition under study (cases) and a very similar group of subjects who do not have the disease or condition (controls). – EMBL (EFO) – MIABIS-2.0-19.
collectionTypeCohort	boolean	1	P	A form of longitudinal study for the analysis of risk factors following a group of people who do not have a disease, and uses correlations to determine the absolute risk of subject contraction. – Wikipedia (rewritten) – MIABIS-2.0-19.
collectionTypeCross-Sectional	boolean	1	P	A type of observational study that involves data collection from a population, or a representative subset, at one specific point in time. – Wikipedia – MIABIS-2.0-19.
collectionType-Longitudinal	boolean	1	P	Research studies involving repeated observations of the same entity over time. In the biobank context, longitudinal studies sample a group of people in a given time period, and study them at intervals by the acquisition and analyses of data and/or samples over time. – P3G – MIABIS-2.0-19.
collectionTypeTwinStudy	boolean	1	P	Twin studies measure the contribution of genetics (as opposed to environment) to a given trait or condition of interest. – MIABIS-2.0-19.
collectionTypeQuality-Control	boolean	1	P	A quality control testing study design type is where some aspect of the experiment is quality controlled for the purposes of quality assurance. – EMBL (EFO) – MIABIS-2.0-19.

Continued on next page...

... continued from previous page.

Attribute	*Type*	*C*	*V*	*Description*
collectionTypePopulation-Based	boolean	1	P	Study done at the population level or among the population groups, generally to find the cause, incidence or spread of the disease or to see the response to the treatment, nutrition or environment. – Wikipedia (rewritten) – MIABIS-2.0-19.
collectionTypeDisease-Specific	boolean	1	P	A collection for which material and information is collected from subjects that have already developed a particular disease. – EMBL (EFO) – MIABIS-2.0-19.
collectionTypeBirth-Cohort	boolean	1	P	A corhort study for which the subjects are followed from the time of birth usually including information about gestation and follow up. – MIABIS-2.0-19.
collectionTypeOther	string	1..n	P	Other type of collection text specified (MIABIS-2.0-19).
collectionOrderOf-Magnitude	integer	1	P	Size of the collection measured as 10^n samples.
Optional				
bioresourceReference	string	0..n	P	Bioresource reference to be cited when the bioresource (biobank/collection) is used for research.
contactIDRef	string	0..n	P	Reference to a contact ID.
contactPriority	integer	0..1	P	Priority of the contact 1..n (i.e., non-negative integer), where the highest priority should be used for contacting about given set of samples. E.g., if a collection has contactPriority=3, the biobank in which the collection resides has contactPriority=10, and the biobankNetwork to which the collection or biobank belongs has contactPriority=7, the biobank contact should be used.
biobankNetworkIDRef	string	0..n	P	Reference to a biobank network ID, to which the collection or biobank belongs; this attribute can also be used for biobank network, where it refers to the superior biobank network).

Continued on next page...

... continued from previous page.

Attribute	*Type*	*C*	*V*	*Description*
geoLatitude	string	0..1	P	Latitude of the biobank in the WGS84 system (the one used by GPS), positive is northern hemisphere.
geoLongitude	string	0..1	P	Longitude of the biobank in the WGS84 system (the one used by GPS), positive is to the East of Greenwich.
collaborationPartners-Commercial	boolean	0..1	P	Biobank/collection can be used for collaboration with commercial partners.
collaborationPartners-Nonforprofit	boolean	0..1	P	Biobank/collection can be used for collaboration with non-for-profit partners.
collectionAcronym	string	0..1	P	Collection acronym according to MIABIS 2.0 – MIABIS-2.0-02.
collectionDescription	string	0..1	P	Collection description according to MIABIS 2.0 – MIABIS-2.0-08.
collectionSexMale	boolean	0..n	P	The sex of the individuals in the sample collection. – MIABIS-2.0-09.
collectionSexFemale	boolean	0..n	P	The sex of the individuals in the sample collection. – MIABIS-2.0-09.
collectionSexUnknown	boolean	0..n	P	The sex of the individuals in the sample collection. – MIABIS-2.0-09.
collectionSex-Undiferrentiated	boolean	0..n	P	The sex of the individuals in the sample collection. – MIABIS-2.0-09.
collectionAgeLow	integer	0..1	P	Age of youngest sample donor at time of sample donation – MIABIS-2.0-10.
collectionAgeHigh	integer	0..1	P	Age of oldest sample donor at time of sample donation – MIABIS-2.0-11.
collectionAgeUnit	string	0..1	P	Unit defining Age Low and Age High. Can be one of the following values: years, months, weeks, days – MIABIS-2.0-08.
collectionAvailable-BiologicalSamples	boolean	0..1	P	Denotes whether biological samples are available (MIABIS-2.0-13).
collectionAvailableSurvey-Data	boolean	0..1	P	Denotes whether survey data are available (MIABIS-2.0-13).

Continued on next page...

... continued from previous page.

Attribute	*Type*	*C*	*V*	*Description*
collectionAvailable-ImagingData	boolean	0..1	P	Denotes whether imaging data are available (MIABIS-2.0-13).
collectionAvailable-MedicalRecords	boolean	0..1	P	Denotes whether medical records are available (MIABIS-2.0-13).
collectionAvailable-NationalRegistries	boolean	0..1	P	Denotes whether register data is associated to the participants in the sample collection/study (MIABIS-2.0-13).
collectionAvailable-GenealogicalRecords	boolean	0..1	P	Denotes whether genealogical records are available (MIABIS-2.0-13).
collectionAvailablePhysio-BiochemMeasurements	boolean	0..1	P	Denotes whether physiological/biochemical measurements are available (MIABIS-2.0-13).
collectionAvailableOther	boolean	0..1	P	Denotes whether other samples/data is available (MIABIS-2.0-13).
temperatureRoom	boolean	0..1	P	Sample storage temperature – Room temperature – SPREC 2.0 (MIABIS-2.0-15).
temperature2to10	boolean	0..1	P	Sample storage temperature – between 2 and 10°C – SPREC 2.0 (MIABIS-2.0-15).
temperature-18to-35	boolean	0..1	P	Sample storage temperature – between -18 and -35°C – SPREC 2.0 (MIABIS-2.0-15).
temperature-60to-85	boolean	0..1	P	Sample storage temperature – between -60 and -85°C – SPREC 2.0 (MIABIS-2.0-15).
temperatureLN	boolean	0..1	P	Sample storage temperature – liquid nitrogen, -150 to -196°C (MIABIS-2.0-15).
temperatureOther	string	0..n	P	Sample storage temperature – other, text specified (MIABIS-2.0-15).

Continued on next page...

... continued from previous page.

Attribute	*Type*	*C*	*V*	*Description*
diagnosisAvailable	string	1..n	P	Diagnosis available in the collection, with the ontology prefix, possibly using * and ? wildcards, and prefix notation to denote diagnosis nomenclature – so far urn:miriam:icd: prefix for ICD-10, and urn:miriam:snomedct: prefix for SNOMED CT (examples being urn:miriam:icd:C*, urn:miriam:snomedct:25*) – MIABIS-2.0-17, adapted.
collectionHeadFirstName	string	0..n	P	First name of a person in charge of the collection.
collectionHeadLastName	string	0..n	P	Last name of a person in charge of the collection.
collectionHeadRole	string	0..n	P	Official role of the person in charge of the collection: typically PI or Director.
collectionSampleAccess-Fee	boolean	0..1	P	Denotes whether access to samples may be obtained on fee-based basis.
collectionSampleAccess-JointProjects	boolean	0..1	P	Denotes whether access to samples may be obtained on joint project basis.
collectionSampleAccess-Description	string	0..n	P	Short description of access rules.
collectionDataAccessFee	boolean	0..1	P	Denotes whether access to data may be obtained on fee-based basis.
collectionDataAccessJoint-Projects	boolean	0..1	P	Denotes whether access to data may be obtained on joint project basis.
collectionDataAccess-Description	string	0..n	P	Short description of access rules.
collectionSampleAccess-URI	string	0..n	P	URI describing access policy for the samples.
collectionDataAccessURI	string	0..n	P	URI describing access policy for the data.
collectionSize	integer	0..1	P	Exact size of the collection to the given date.
collectionSizeTimestamp	integer	0..1	P	Date to which the size of the collection was valid, absolute time in ISO 8601 format.

D. Data Provider Agreement Template (Biobank – BBMRI-ERIC)

This appendix provides a template of the *an agreement between the biobank (provider of the data) and BBMRI-ERIC*. The template is based on the results of BioMedBridges project.[25] The data provider agreement is a specific form of DTA, focused on providing the data into a common data set.

[25] Funded by the European Commission by FP7 under grant number 284209 (BioMedBridges).

Provider Agreement for Personal Data

Preamble

This agreement governs the exchange of personal data within the CRC-Cohort developed within the ADOPT project (this project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550). All members of the project are bound by the Ethical and Governance Framework of ADOPT which forms the basis of this agreement. The parties of this agreement acknowledge the validity and applicability of Ethical and Governance Framework in its current version.

1. Parties

The undersigned, the Institution X [fill in official name of legal entity that is authorized to enter into this agreement], a [fill in type of legal entity, e.g. foundation, charitable trust, corporation (ltd. Inc.)], incorporated, organized and duly existing under the laws of the [fill in appropriate jurisdiction], with its principal office at [insert address], hereby legally represented by [insert name of legal representative],

and

BBMRI-ERIC, a Biobanking and Biomolecular Resources Research Infrastructure-European Research Infrastructure Consortium incorporated, organized and duly existing under the laws of the European Union and Austria, with its principal office at Neue Stiftingtalstraße 2/B/6, 8010 Graz, Austria, hereby legally represented by Erik Steinfelder, Director General of BBMRI-ERIC.

have agreed to be bound by the provisions set out in this Agreement.

Whereas, the Institution X is a [e.g. population/clinical biobank] established with the aim to facilitate research on its collection of [e.g. human biological samples]; BBMRI-ERIC is willing to cooperate with Institution X in building a data base called Colon Cancer Cohort in order to enhance research on data.

2. Scope of Exchange

The Institution X provides to BBMRI-ERIC the following data: NN [e.g., 300] colorectal cancer cases compliant with the inclusion criteria and data in the structure prescribed in Section B.1

of Annex I to this contract (see Section 13). The data has to be usable for the purpose defined in the Section 1 of the Annex I.

The data have anyway to meet the following standards:

- Data has to be entered into a software application provided by BBMRI-ERIC, or handed over to BBMRI-ERIC in an agreed-upon, documented machine readable format.
- The Institution X is required to keep documentation on how the data was collected for BBMRI-ERIC (Section 2.5 of Annex I) and make it available upon request if reproducibility concerns are raised.

3. Data Protection

The Institution X confirms that for the purposes of this Agreement it is entitled to transfer the personal data to BBMRI-ERIC and that consent covering the intended use has been obtained from the relevant donors/data subjects.

The Institution X and BBMRI-ERIC will retain the data in a secure network system at such standard as would be reasonably expected for the storage of valuable and proprietary sensitive/confidential data. BBMRI-ERIC shall refrain from tracing or identifying the identity of any donors who provided the samples to Institution X. BBMRI-ERIC agrees to preserve, at all times, the confidentiality of information pertaining to identifiable donors. BBMRI-ERIC agrees to give access to the data, in whole or part, to any third party only after conducting the access procedure defined in Annex I. BBMRI-ERIC shall limit access to and processing of the data to those employees or other authorized representatives of BBMRI-ERIC who: (i) need to process such data in order to conduct their work in connection with the data and (ii) have signed agreements with BBMRI-ERIC obligating them to maintain the confidentiality of the data and any information to be derived thereof or disclosed to them.

The BBMRI-ERIC shall not attempt to contact any donor/data subject. Should there be a need to contact donors/data subjects, such as in case of results or incidental findings, the contact requests will be handed over to the Institution X by BBMRI-ERIC.

Institution X and BBMRI-ERIC shall take reasonable steps to delete all personal data for a given subject when the Institution deems that subject to have withdrawn his or her consent. BBMRI-ERIC confirms that it will deal promptly and appropriately with any withdrawals by donors/data subjects which the Institution X notify to BBMRI-ERIC.

On the termination of this agreement, BBMRI-ERIC shall delete all personal data and confirm to the other Institution (in writing) that this has taken place.

Any provisions of this agreement intended to protect the rights of human donors/data subjects shall survive the expiry or termination of this agreement.

4. Intellectual Property

Handling of intellectual property will be covered by BBMRI-ERIC in the Data Transfer Agreement between BBMRI-ERIC and a third party (subject to the access procedure defined in Annex I) in a way that (a) both, Institution X and BBMRI-ERIC, shall be entitled to any inventions to the extent that these result from their own independent use of the data; (b) each Institution (Institution X, BBMRI-ERIC, third party) shall grant the other Institution a worldwide non-exclusive royalty-free irrevocable research licence with respect to any such inventions; (c) if one Institution elects not to seek any intellectual property protection with respect to such inventions he shall transfer any such rights to the other Institution at no cost.

5. Return and publication of Results

The Institution X and BBMRI-ERIC agree that both institutions will publish - not before the date of publication of a paper that describes the results of any analyses of the data - via the each Institution's website:

- General information about the analysis to the public.

BBMRI-ERIC will require free-of-charge return of data from any third party accessing the data (subject to the access procedure defined in Annex I). Returned data will be offered free of charge by BBMRI-ERIC to the Institution X for integration into their data sets.

6. Credits

The Institution X and BBMRI-ERIC agree to acknowledge the source of the data in any publications or other public disclosures reporting use of it. BBMRI-ERIC will also require any third party accessing the data (subject to access procedure defined in Annex I) to use the same acknowledgement in their publications or other public disclosures reporting the use of the data.

The following form of words should be used: "This project has been supported by CRC-Cohort developed in ADOPT BBMRI-ERIC project (this project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550). Institution X has provided data for this project.".

7. Reports/Notification

BBMRI-ERIC shall provide a copy of any report of its Results that derive from use of the Resource to the Institution X in any format (e.g., journal or conference paper, on-line report).

8. Expiry/Termination

This agreement is perpetual, unless terminated by the mutual written agreement of the parties.

Upon termination of this Agreement:

- The grant of rights to the Institution X and BBMRI-ERIC will be automatically terminated;
- BBMRI-ERIC shall destroy the Data or otherwise render it inaccessible.

9. Limitation of Liability and Indemnity

The Institution X and BBMRI-ERIC will indemnify each other against all losses (whether direct or indirect, reasonably foreseeable or specifically contemplated by the parties), damages, costs, expenses (including but not limited to reasonable legal costs and expenses) that it incurs as a result of: (i) the use, storage or disposal of human personal data of the other Institution; or (ii) any negligence or willful default of the other Institution, provided that the Institution agrees to use its reasonable endeavours to mitigate any loss.

10. Force majeure

If any party is prevented from, hindered or delayed in performing any of its obligations under this Agreement by reason of a Force Majeure Event, such party will promptly notify the other of the date of its commencement and the effects of the Force Majeure Event on its ability to perform its obligations under this Agreement. If mutually agreed by the parties, then the obligations of the party so affected will thereupon be suspended for so long as the Force Majeure Event may continue. The party affected by a Force Majeure Event will not be liable for any failure to perform such of its obligations as are prevented by the Force Majeure Event provided that such party will use every reasonable effort to minimise the effects thereof and will resume performance as soon as possible after the removal of such Force Majeure Event. If the period of non-performance exceeds 28 days from the start of the Force Majeure Event then the non-affected party will have the option, by written notice to the other party, to terminate this Agreement. For the purpose of this clause, Force Majeure Event means any event beyond the reasonable control of a party including, without limitation, acts of God, war, terrorism, riot, civil commotion, malicious damage, compliance with any law or governmental order, rule, regulation or direction, accident, fire, flood or storm. For the avoidance of doubt, strike, industrial action, failure of technology systems, third party insolvency and failure of the Institution X and/or BBMRI-ERIC or any other third party will not be considered to be Force Majeure Events. The provisions of this clause will not affect any other right which either party may have to terminate this Agreement.

11. Applicable law and jurisdiction

This Agreement will be governed by and construed in accordance with the laws of Austria; parties agree that the Austrian courts will have exclusive jurisdiction over any suit, action, proceedings or dispute arising out of, or in connection with, this Agreement.

12. General

This Agreement governs the relationship between the parties to the exclusion of any other terms and conditions and, together with any other document referred to in this Agreement, constitutes the whole agreement between the parties in relation to the subject matter hereof.

If there is any conflict between the provisions of this Agreement and any of the annexes and related documents (including, but without limitation, the provisions of the Access Procedures) then the provisions of this Agreement will apply.

A waiver, delay or forbearance by either party, whether express or implied, in enforcing or exercising any of its rights or remedies hereunder will not constitute a waiver of such right or remedy.

No provision of this Agreement is intended to be enforceable by any person who is not a party to this Agreement and nor are any rights granted to any third party under statute or otherwise.

All variations to this Agreement must be agreed, set out in writing and signed on behalf of the parties before they take effect.

13. Attachments

This Agreement incorporates the following Annexes:

- CRC-Cohort Data Protection Policy (Annex I).

Signatures

Yours faithfully Accepted and agreed

For and on behalf of the Institution X

For and on behalf of BBMRI-ERIC

E. Data Transfer Agreement (DTA) Template (BBMRI-ERIC–Researcher)

This appendix provides a template of the *an agreement between the researcher (requester of the data) and BBMRI-ERIC*. The template is based on the results of BioMedBridges project.[26]

[26] Funded by the European Commission by FP7 under grant number 284209 (BioMedBridges).

Data Transfer Agreement (DTA)

Preamble

This agreement governs the transfer of human personal data. It is designed for cases, where no cooperation agreement exists between the contractors.

1. Parties

The undersigned, BBMRI-ERIC, a Biobanking and Biomolecular Resources Research Infrastructure-European Research Infrastructure Consortium incorporated, organized and duly existing under the laws of the European Union and Austria, with its principal office at Neue Stiftingtalstraße 2/B/6, 8010 Graz, Austria, hereby legally represented by Erik Steinfelder, Director General of BBMRI-ERIC,

and

Institution Y [fill in official name of legal entity that is authorized to enter into this agreement], a [fill in type of legal entity, e.g. foundation, charitable trust, corporation (ltd. Inc.)] incorporated, organized and duly existing under the laws of the [fill in appropriate jurisdiction], with its principal office at [insert address], hereby legally represented by [insert name of legal representative],

and/or

Company Z [fill in official name of legal entity that is authorized to enter into this agreement], a [fill in type of legal entity, e.g. corporation (ltd., inc.)], incorporated, organized and duly existing under the laws under the laws of the [fill in appropriate jurisdiction], with its principal office at [insert address], hereby legally represented by [insert name of legal representative], and/or [..]

[Add any other party which is to be a party to the DTA],

Whereas, BBMRI-ERIC is an international research infrastructure consortium established with the aim to facilitate research on human biological material and data; Institution Y (hereinafter: Recipient) is a [for profit or non-profit] research institute willing to conduct research on certain Data from BBMRI-ERIC; BBMRI-ERIC is willing to transfer certain Data to the Recipient;

have agreed to be bound by the provisions set out in this Agreement.

2. Scope of Supply

BBMRI-ERIC provides to the Recipient the following data: … [fill in precisely or point to an attachment where the provision is described].

The Recipient acknowledges that the data are provided on an "as is" basis without any warranty of satisfactory quality or fitness for a particular purpose or use or any other warranty, express or implied.

3. Data Protection

Annex A summarises the data and/or data that BBMRI-ERIC will make available to the Recipient in accordance with their approved Application [reference number]. The timeframe and methodology by which the data and/or data will be dispatched is also set out in Annex A.

BBMRI-ERIC confirms that for the purposes of this DTA it is entitled to supply the data/and or personal data to the Recipient and that consent covering the intended use has been obtained from the relevant donors/data subjects.

The Recipient will use data for purposes of the analyses set forth and within the limits set by the Research protocol only. The Recipient confirms that the Approved Research Project has been subject to independent scientific review by a recognised body in the manner described in the Application and that the planned use of the data/data has approval of the appropriate ethics and scientific committees. The Recipient confirms that all work using the data will be carried out in compliance with all applicable laws, regulations, guidelines and approvals.

The Recipient will retain the data in a secure network system at such standard as would be reasonably expected for the storage of valuable and proprietary for sensitive/confidential data. The Recipient shall refrain from tracing or identifying the identity of any donors who provided the data. Recipient agrees to preserve, at all times, the confidentiality of information pertaining to identifiable donors. The Recipient agrees not to give access to data, in whole or part, or any identifiable data derived from the data, to any third party. The Recipient shall limit access to and processing of the data to those employees or other authorized representatives of Recipient who: (i) need to process such data in order to conduct their work in connection with the data and the Protocol and (ii) have signed agreements with the Recipient obligating them to maintain the confidentiality of the data and any information to be derived thereof or disclosed to them.

The Recipient shall not attempt to contact any data subject.

Recipient shall take reasonable steps to delete data for a given subject when BBMRI-ERIC deems that subject to have withdrawn his or her consent. The Recipient confirms that it will deal promptly and appropriately with any withdrawals by donors/data subjects which BBMRI-ERIC notify to the Recipient.

On the Completion of the Research Project or on the termination of this agreement, the Recipient will delete the data and confirm to BBMRI-ERIC (in writing) that this has taken place.

On reasonable notice to the Recipient, and in order to confirm or investigate compliance with the provisions of this DTA, BBMRI-ERIC may itself or via appropriate third parties:

- choose to inspect the premises and other relevant facilities of the Recipient, in order to review the security, storage or other arrangements for the data;
- request such additional information about the Approved Research Project and/or its progress as BBMRI-ERIC may, from time to time, reasonably require.
- BBMRI-ERIC will bear the costs of such audits unless a data default within the procedures and processes of the Recipient is discovered, in which case the Recipient will be obliged to reimburse the reasonable costs of BBMRI-ERIC and any relevant third parties.

Any provisions of this agreement intended to protect the rights of human donors/data subjects shall survive the expiry or termination of this agreement.

4. Intellectual Property

Title to the data is and remains in the ownership of BBMRI-ERIC and the data are made available to the Recipient as a service to the research community.

The Recipient shall be entitled to any inventions to the extent that these result from his own independent use of the data. He shall grant BBMRI-ERIC a worldwide non-exclusive royalty-free irrevocable research licence with respect to any such inventions. BBMRI-ERIC is entitled to grant this license on the same terms to BBMRI-ERIC partner biobanks that contributed to the data delivered to the Recipient: [list the biobanks together with their juridical persons for clarity reasons]. If the Recipient elects not to seek any intellectual property protection with respect to such inventions he shall transfer any such rights to BBMRI-ERIC at no cost.

To the extent that BBMRI-ERIC, BBMRI-ERIC partner biobanks that contributed to the data delivered, and the Recipient have each contributed to an invention with respect to the data, they shall jointly own any rights to such an invention. Inventions made solely by the employees or agents of one party shall be owned by that party.

Except as expressly set forth in this Section, nothing herein shall be deemed to grant to either BBMRI-ERIC or BBMRI-ERIC partner biobanks that contributed to the data delivered or Recipient any rights under the other party's patents, patent applications, trademarks, copyrights, trade secrets, know how (whether patentable or unpatentable) or other intellectual property rights.

5. Return, publication, and retention of Results

The Recipient agrees that BBMRI-ERIC will publish - at a time not before the date of publication of a paper that describes the results of any analyses of the Material - via BBMRI-ERIC's website:

- General information about the analysis to the public.
- Summary data about the results to registered users of BBMRI-ERIC's website.
- De-identified subject-specific data about the results to registered users of BBMRI-ERIC's website.

The Recipient must retain the obtained research data for minimum of 10 years and make it findable, accessible, interoperable and reusable (FAIR) in a data protection respecting manner to enable reproducibility and verifiability of findings.

6. Credits

The Recipient agrees to acknowledge the source of the data in any publications or other public disclosures reporting use of it. The following form of words should be used: "This project has been supported by CRC-Cohort developed in ADOPT BBMRI-ERIC project (this project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550). [List of contributing BBMRI-ERIC partner biobanks] have provided data for this project.".

7. Reports/Notification

The Recipient shall provide a copy of any report of its Results that derive from use of the Resource to BBMRI-ERIC in any format (e.g. paper journal, on-line report, meeting abstract).

Notices required under this DTA will be in writing and will be delivered by email to the addresses set out below or (in the event of a failure to deliver an email) by post to BBMRI-ERIC or the Recipient and will be deemed to be given, in the case of delivery by email, upon receipt at the Recipient 's email server (unless an automatic response indicating an undeliverable message is received) and, in the case of delivery by post, on the date of delivery (or, if not a business day, on the first business day thereafter).

8. Expiry/Termination

This agreement shall expire …[fill in date], unless earlier terminated by the mutual written agreement of the parties.

BBMRI-ERIC will be entitled to terminate this DTA forthwith by written notice to the Recipient if:

- The Recipient commits any breach of a data provision of this DTA and, in the case of a breach capable of remedy, fails to remedy the same within 20 days after receipt of a written

notice giving particulars of the breach and requiring it to be remedied; a breach will be considered capable of remedy if the Recipient can comply with the provision in question in all respects other than as to the time of performance, provided that time of performance is not of the essence.

- The Recipient PI ceases to be employed (or otherwise engaged by) the Recipient Institution; or
- The Recipient Institution ceases, is likely to cease, or threatens to cease carrying on business.

The rights to terminate this DTA given by this clause will be without prejudice to any other right or remedy of either party in respect of the breach concerned, if any, or any other breach.

Upon expiry or termination of this Agreement:

- The grant of rights to the Recipient will be automatically terminated;
- The Recipient shall delete the data.

9. Charges/Payment

In consideration for BBMRI-ERIC's entering into this Agreement, the Recipient agrees to pay BBMRI-ERIC an amount of [specify fee and VAT, if applicable] by wire transfer to BBMRI-ERIC, account Number [..], BIC/swiftcode [..].

This DTA is conditional on the Access Charges being paid and so, for the avoidance of doubt, no biodata/data will be provided to the Recipient until or unless the access charges are received in full.

10. Assignment and sub-contracting

Neither party will be entitled to assign this DTA or any of its rights or obligations hereunder without first having received the written approval of the other party, which approval not to be unreasonably withheld or delayed. The Recipient will not sub-contract the performance of any of its obligations under the DTA or any part thereof without having first obtained the prior written consent of BBMRI-ERIC, such consent not be unreasonably withheld. In the event that consent is granted, the Recipient shall be responsible for the acts, defaults and omissions of its sub-contractors as if they were the Recipient's own, and any consent given will not relieve the Recipient of any of its obligations under this DTA.

11. Limitation of Liability and Indemnity

The Recipient will indemnify BBMRI-ERIC against all losses (whether direct or indirect, reasonably foreseeable or specifically contemplated by the parties), damages, costs, expenses (including but not limited to reasonable legal costs and expenses) that it incurs as a result of: (i) the use, storage or disposal of human personal data by the Recipient; or (ii) any negligence or wilful default of the Recipient, provided that BBMRI-ERIC agrees to use its reasonable endeavours to mitigate any loss.

12. Force majeure

If any party is prevented from, hindered or delayed in performing any of its obligations under this DTA by reason of a Force Majeure Event, such party will promptly notify the other of the date of its commencement and the effects of the Force Majeure Event on its ability to perform its obligations under this DTA. If mutually agreed by the parties, then the obligations of the party so affected will thereupon be suspended for so long as the Force Majeure Event may continue. The party affected by a Force Majeure Event will not be liable for any failure to perform such of its obligations as are prevented by the Force Majeure Event provided that such party will use every reasonable effort to minimise the effects thereof and will resume performance as soon as possible after the removal of such Force Majeure Event. If the period of non-performance exceeds 28 days from the start of the Force Majeure Event then the non-affected party will have the option, by written notice to the other party, to terminate this DTA. For the purpose of this clause, Force Majeure Event means any event beyond the reasonable control of a party including, without limitation, acts of God, war, terrorism, riot, civil commotion, malicious damage, compliance with any law or governmental order, rule, regulation or direction, accident, fire, flood or storm. For the avoidance of doubt, strike, industrial action, failure of technology systems, third party insolvency and failure of BBMRI-ERIC or any other third party will not be considered to be Force Majeure Events. The provisions of this clause will not affect any other right which either party may have to terminate this DTA.

13. Applicable law and jurisdiction

This DTA will be governed by and construed in accordance with the laws of Austria; parties agree that the Austrian courts will have exclusive jurisdiction over any suit, action, proceedings or dispute arising out of, or in connection with, this Agreement.

14. General

This DTA governs the relationship between the parties to the exclusion of any other terms and conditions and, together with any other document referred to in this Agreement, constitutes the whole agreement between the parties in relation to the subject matter hereof.

If there is any conflict between the provisions of this DTA and any of the annexes and related documents (including, but without limitation, the provisions of the Access Procedures) then the provisions of this DTA will apply.

A waiver, delay or forbearance by either party, whether express or implied, in enforcing or exercising any of its rights or remedies hereunder will not constitute a waiver of such right or remedy.

No provision of this DTA is intended to be enforceable by any person who is not a party to this Agreement and nor are any rights granted to any third party under statute or otherwise.

Nothing in this DTA will create a partnership, joint venture or relationship of agency between the parties.

All variations to this DTA must be agreed, set out in writing and signed on behalf of the parties before they take effect.

15. Attachments

This Agreement incorporates the attached terms and conditions (including any documents and/or data that are referred to in them), the Annexes and where applicable the contents of the Preliminary and Main Application Forms [reference number].

Signatures

Yours faithfully Accepted and agreed

For and on behalf of BBMRI-ERIC For and on behalf of Recipient Institution

For and on behalf of the Recipient Principal Investigator.

F. Acceptable Use Policy of BBMRI-ERIC Services

The Acceptable Use Policy (AUP) specifies conditions that must be accepted by any authenticated User of any BBMRI-ERIC service dealing with personal data or anonymized data (for example, but not limited to, BBMRI-ERIC Locator/Finder and BBMRI-ERIC Negotiator).

BBMRI-ERIC
Biobanking and
BioMolecular resources
Research Infrastructure

Acceptable Use Policy of BBMRI-ERIC Services

Version 1.0, 2017-08-23

1 Preamble

These usage conditions must be accepted by any authenticated User of any BBMRI-ERIC Service dealing with personal data or anonymized data (for example, but not limited to, BBMRI-ERIC Locator and BBMRI-ERIC Negotiator).

2 Definitions

BBMRI-ERIC Locator A service for searching preliminary availability information on samples and data sets. To be available in 2018.

BBMRI-ERIC Negotiator A service for facilitating access to BBMRI-ERIC partner biobanks, by orchestrating and simplifying the communications between researchers (requesters) and biobankers. `https://negotiator.bbmri-eric.eu/`.

Contributing Biobank Biobank which contributed the data or biological material via BBMRI-ERIC to the researcher.

Research Participant Patient or donor, who has consented to and contributed her/his data or biological material to be used in research.

3 Common Conditions of Use

- User agrees to be a bona fide researcher with (1) an intention to generate new knowledge and understanding using rigorous scientific methods, (2) an intention to publish the research findings and share the derived data in the scientific community, without restrictions and with minimal delay, for wider scientific and eventual public benefit, and where (3) the intended activities are not inconsistent with legal and ethical requirements or widely recognised good research practice.
- The User will avoid any attempts to reverse privacy enhancing technologies (i.e., pseudonymization, anonymization) applied to the data and/or to (re-)identify individual Research Participants contributing the data and/or biological.

- If possible,[1] any incidental findings will be reported back to the Contributing Biobanks, which in turn process the incidental findings based on their policy and based on the conditions specified in the informed consent.

4 Service-Specific Conditions of Use

- *BBMRI-ERIC Negotiator:* The User agrees to honor the BBMRI-ERIC Harmonized Access Procedure,[2] including data protection and confidentiality principles.

[1] Application of privacy-enhancing technologies may render reporting incidental findings back impossible, e.g., when found on anonymized data sets.

[2] https://doi.org/10.5281/zenodo.823013

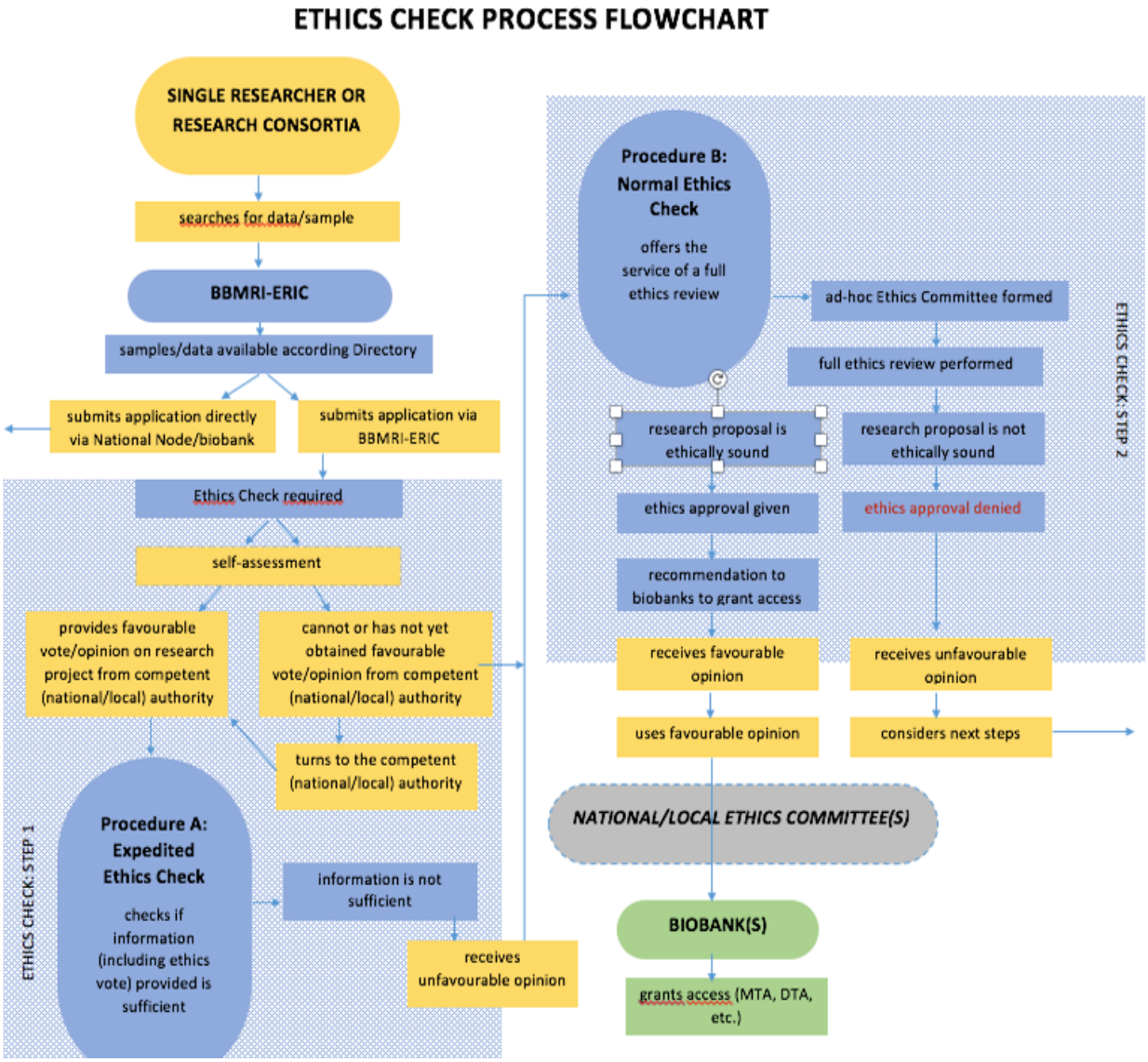
Document Log

Version 1.1, 2017-08-24 Added Service-specific policy for Negotiator. Renamed from Terms and Conditions to Acceptable Use Policy. Added Document Log. Author: Petr Holub

Version 1.0, 2017-07-12 Initial version. Author: Petr Holub

G. Ethics Check

Based on the Ethics Check principles in [3], the following practical Ethics Check process has been developed for BBMRI-ERIC:



H. BBMRI-ERIC Access Policy

This is a temporary appendix with the BBMRI-ERIC Access Policy [10], which is expected to be approved by the BBMRI-ERIC Assembly of Members in Fall 2017.

BBMRI-ERIC
Biobanking and
BioMolecular resources
Research Infrastructure

BBMRI-ERIC Policy
for
Access to and Sharing of Biological Samples and Data

The Statutory Basis for a BBMRI-ERIC Access and Sharing Policy

Having regard to the statutes of BBMRI-ERIC, stating that the Assembly of Members shall:

- *adopt all rules, regulations and policies necessary for the sound management of the Work Programme, in particular the access procedure to biological resources, data in Biobanks and services developed by BBMRI.* (Article 10.6.c).

Furthermore, that BBMRI-ERIC shall:

- *make samples and data in databases affiliated with or developed by BBMRI-ERIC Partner Biobanks available to researchers and research institutions according to the access procedure and criteria as approved by the Assembly of Members. Access shall respect conditions set by sample and data providers that affiliate their databases to BBMRI-ERIC. No provision in these Statutes should be understood as seeking to restrict the right of owners of Biobanks or Biomolecular Resources affiliated with BBMRI-ERIC to decide on providing access to any samples and data* (Article 18.1),
- *shall provide access to samples and related clinical data based on the scientific excellence of the proposed project as determined by an independent peer review and after ethical review of the research project proposal* (Article 18.2),
- *that BBMRI-ERIC shall seek to ensure that the source of samples and data be appropriately acknowledged and shall request that such attribution be maintained in subsequent use of the samples and data* (Article 18.3).

Having regard to *The European Research Infrastructure for BioBanking and Biomolecular Resources Partner Charter* as approved by the Assembly of Members on 29 April 2015, stating the agreed principles for access policy, data protection and management policy and informed consent,

The following *BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data*, based on state-of-the-art ethical and legal premises for such an access policy and dedicated work done by the Common Services ELSI and IT as well as the EU Project ADOPT BBMRI-ERIC's Work Package 5, is proposed. Once approved by the Assembly of Members it should be used by all National Nodes of BBMRI-ERIC and the partner biobanks when partaking in exchange of samples and data as well as in all communication and training activities pertaining to their national biobank activities.

BBMRI-ERIC
Biobanking and
BioMolecular resources
Research Infrastructure

BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data

Introduction

BBMRI-ERIC[1] is a pan-European research infrastructure which facilitates access to human biological samples (e.g., tissue, blood, DNA) and associated clinical and research data from individual biobanks. BBMRI-ERIC is a European infrastructure with the aim to encourage and expedite effective and ethical access to samples and data from biobanks, preferably in the context of high-level research collaboration between providers (e.g., biobanks, including scientists and physicians who contributed to the biobanks) and requesters. As of early 2017, BBMRI-ERIC consists of 19 Member States and one international organisation (IARC). BBMRI-ERIC operates on a non-economic basis. This access policy presents three areas of guidance: i) ethical principles; ii) governance procedures; and iii) practical procedures for access. Together with existing legal frameworks, these three areas provide the ethical and legal framework and practical procedures to guide the access to, and use of, biological samples and associated data as well as to tools and resources developed by BBMRI-ERIC.

This policy is a binding document for BBMRI-ERIC itself, for BBMRI-ERIC partner biobanks, and for any requesters, who are seeking access to samples/data from BBMRI-ERIC partner biobanks via BBMRI-ERIC. It will not supersede access policies and procedures of individual biobanks; but will provide a framework that BBMRI-ERIC partner biobanks must adhere to.

This policy will be amended/changed in accordance with changes in the regulatory framework.

1. Legal Premise

All proceedings related to access and sharing must be compliant with national and European legislation, e.g., the EU General Data Protection Regulation with its Code of Conduct, both coming into effect as of 28 May 2018. It is the national legislation applying to the data controller/person responsible for a biorepository, a registry, or a collection of personal data that, in turn, applies to the processing of samples and data, irrespective of where the samples and data are used. The processing of data and the use of biological samples must be compliant with the provisions of the informed consent form and/or decision of an ethical review board, and/or a data protection authority/officer if applicable. If none of the above is applicable, it must be compliant with national legislation.

2. Governing Ethical Principles

BBMRI-ERIC will facilitate and support access to samples and data from participating biobanks, as well as access to, and the use of, resources and tools developed by BBMRI-ERIC according to the following principles:

I. **Scientific integrity**: BBMRI-ERIC requesters and providers are expected to act in an honest, transparent, equitable manner and uphold the highest standards of quality in scientific research.

II. **Responsibility and accountability**: It is both the requesters' and providers' responsibility to ensure that they have read and understood the relevant policies and procedures and that they act in accordance with them (including the biobank's as well as BBMRI-ERIC's policies and procedures). Should policies or procedures established by BBMRI-ERIC be contravened, the requester or provider is expected to report this to BBMRI-ERIC immediately.

III. **Respect for responsible governance regarding data and research:** Requesters and providers are expected to take the necessary precautions and safeguards to avoid subjects' privacy breaches. This entails protecting their personal data and putting in place state-of-the-art safety measures for data security.

IV. **Respectful use of limited resources**: Request for access to biological samples of a limited nature will be particularly parsimonious. Requesters are expected to request only as much as is required and for results that cannot be effectively achieved otherwise.

V. **Accessibility to research results**: Requesters should be willing to make their research results accessible for academic purposes on a royalty-free basis and in a timely manner.

VI. **Attribution**: The intellectual investment of investigators involved in the creation of data registries and bio-repositories is often substantial, and should be acknowledged. This should be specified in Material and Data Transfer Agreements (MTA/DTA) signed by both parties.

VII. **Respect for intellectual property**: Sharing of data and biological samples needs to be performed in a way that protects intellectual property rights of the parties involved . It also needs to address the requirements of institutions and third-party funders.

VIII. **Equity and inclusivity of users**: *Bona fide* researchers who meet the relevant criteria should be granted access based on fair and non-discriminatory terms.

IX. **Public engagement**: BBMRI-ERIC supports the engagement of relevant stakeholders and the public and welcomes their active participation in biobanking.

X. **Reciprocity**: Stewardship also implies giving something back. Feedback regarding general results should be channelled towards institutions and patients.

XI. **Confidentiality:** BBMRI-ERIC and its partner biobanks shall treat all the access requests confidentially and will not use them for any purpose other than assessing the availability of the samples/data and access provisions.

3. Procedures Governing Access to and Use of Samples/Data

I. BBMRI-ERIC encourages BBMRI-ERIC partner biobanks to refer to the *European Charter for Access to Research Infrastructures*[2] *(ECfARI)* and the *International Charter of Principles for Sharing Biological Samples and Data*[3] when updating existing access policies or defining new ones. Due to the diversity of the BBMRI-ERIC partner biobanks, access units and modes are to be defined by each biobank. The ECfARI requires research infrastructure to define their access units and access modes. The '*access units*' [ECfARI 3.d.] that BBMRI-ERIC recommends are samples and resources consumed for preparation and delivery of data sets. BBMRI-ERIC recommends that '*access mode*' [ECfARI 5.b.] is excellence-driven access [ECfARI 5.b.1.]. Biobanks can adjust these recommendations according to their preferences.

II. The samples/data remain under the stewardship of the BBMRI-ERIC partner biobanks as the original source, unless otherwise specified under a separate agreement. Consequently, BBMRI-ERIC only facilitates access, while BBMRI-ERIC partner biobanks actually grant and provide access. Access should be based on requests for specified research projects.

III. The quality of data and biological samples shall be ensured by the provider.

IV. The requester needs to ascertain that the samples and data provided are stored in a secure storage and operation facility accompanied by an appropriate access policy, including a description of who can access the facility, the time-period the samples/data will or need to be stored, and concrete steps for sharing samples/data.

V. Material Transfer Agreements (MTAs), Data Transfer Agreements (DTAs) or Data Access Agreements (DAAs) should always be used to govern material transfer between parties.

VI. Samples/data can only be used for academic or industrial research purposes, depending on the legislation of the Member State or international organisation: The usage and limitations need to be specified in an MTA/DTA between the requester and the BBMRI-ERIC partner biobank.

VII. All projects using human biological materials and derived data (beyond the original project for which samples/data were initially collected and provided) are subject to the overarching principle above, i.e., they must also be evaluated by an appropriate and legitimate ethical review board.

VIII. The entity which provides biological samples or personal data shall explicitly document any restriction of use or obligation applicable to these biological samples or data (e.g., the limited scope of purpose imposed by the consent form, the obligation to report incidental findings, publication limits such as non-discrimination clauses, etc.).

IX. Requests for access to samples/data issued by requesters will be required to follow the request procedure for samples/data via the BBMRI-ERIC IT services below.

X. DTAs or DAAs should be used for parties outside the EU, unless there are specific legal provisions between the EU and the third country (e.g., Privacy Shield for U.S., or countries accepted by the European Commission in accordance with Article 25, Directive 95/46/EC) using standard EC contract clauses, e.g., Commission Decision 2001/497/EC, C(2004)5721.

XI. In order to maximise the value of the biobank resources, providers may request that provenance data as well as data derived from samples/data are transferred back to the respective provider free of charge (so-called 'return of data'). If the provider does not have the capacity to store the data, the provider may contact BBMRI-ERIC to facilitate storage.

XII. Access will be cost-neutral for BBMRI-ERIC partner biobanks. It may be that BBMRI-ERIC partner biobanks require the requesters to partially or fully cover the costs incurred in providing samples and/or data. Cost aspects must be regulated in the MTA/DTA between the requester and the

BBMRI-ERIC partner biobank.

4. Request Procedure for Access to Samples/Data via BBMRI-ERIC IT Services

The basic framework governing the request procedure for accessing samples/data via the BBMRI-ERIC IT services comprises the following steps:

Step-1 *Registration of 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 IT services. Each request must include information about the approved/proposed research project including its ethical approval status, expected properties and amount of samples/data and their anticipated use as well as the destination of the samples (if different from the location of the requester). A provider may either request refinement of the request or provide *Availability Information* to the requester via BBMRI-ERIC. In compliance with the governing ethical principles in paragraph XI., *Availability Information* is treated as confidential by BBMRI-ERIC, i.e., it will not be disclosed to other providers. Providers will not use requests for any other purpose than assessing the availability of the requested samples/data and providing offers.

Step-3 *Access control & samples/data delivery:* After receiving adequate *Availability Information*, the requester follows up directly with the provider (biobank) in order to provide any additional information needed for assessing whether access can be granted. As part of this process, the provider must comply with the regulatory and ethical conditions (e.g., data protection regulations, assessment of compliance of the informed consent with the approved/proposed project, check whether the amount of deployable/extraditable samples required is scientifically justified) and transfer liability to the requester by using MTAs/DTAs as deemed appropriate. The provider has to decide whether samples/data are released for the project requested. Similarly, access to deliverable/extraditable samples may be subject to prioritisation. For approved requests, the MTA/DTA will need to be executed and access charges paid before samples/data are released to the requester.

Step-4 *Return of results:* Providers need to collect reports on project outcomes for accountability purposes regarding the utilisation of BBMRI-ERIC infrastructure. Providers are encouraged to require the return of derived data from the requester (see *Access Criteria* above) and integrate this requirement into their biobank policy and the respective MTA/DTA.

Step-5 *Request completion notification:* For each request obtained via BBMRI-ERIC, for which *Availability Information* was provided according to Step-2 and for which Step-3 has been completed, the BBMRI-ERIC partner biobanks are required to inform BBMRI-ERIC whether the request has been completed successfully (whether samples and/or data were provided to the requester and/or whether Step-4 and -5 were also completed), 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.

As access facilitator, BBMRI-ERIC provides infrastructure implementing Step-1, Step-2 and Step-5. BBMRI-ERIC is not directly involved in Step-3 and Step-4.

5. Implementation of the BBMRI-ERIC Access Procedure

Step-1, Step-2 and Step-5 of the access procedure will be implemented through a tool called BBMRI-ERIC *Negotiator*[4], to be delivered by the BBMRI-ERIC Common Service IT. Step-1 will use a federated authentication mechanism in conjunction with the *Negotiator*.

6. Definition of Terms

Samples/data

Biological samples or data stored in, and under stewardship of, one of the BBMRI-ERIC partner biobanks, as well as data derived through the use of the requestor.

BBMRI-ERIC partner biobank

Biobanks participating in BBMRI-ERIC infrastructure as part of the National/Organisational Nodes that have signed the BBMRI-ERIC *Partner Charter*. Please note that individual biobanks remain in control over ultimately granting/denying access to potential users/requesters.

Requester

A qualified person requesting samples/data. Needs to be registered as specified in Step-1 and -2.

Bona fide researcher

A researcher with:

1. an intention to generate new knowledge and understanding using rigorous scientific methods,
2. an intention to publish the research findings and share the derived data in the scientific community, without restrictions and with minimal delay, for wider scientific and eventual public benefit, and where
3. the intended activities are not inconsistent with legal and ethical requirements or widely recognised good research practice
4. they have a *bona fide* research project

***Bona fide* research project:** In practical terms, a research project or proposal that has been approved by a recognised funder, or a researcher that belongs to a research organisation that has the capability to lead or participate in high quality, ethical research should normally be considered *bona fide*.

National/Organisational Node

A National Node or an Organisational Node as defined in the Statutes of BBMRI-ERIC (Article 1.6 and 1.8).

Provider

A BBMRI-ERIC partner biobank providing samples/data.

MTA

A contract between the requester and the partner biobank specifying conditions under which the biological material and/or data are transferred from the biobank to a recipient. A data-only transfer agreement is sometimes called a Data Transfer Agreement (DTA) or Data Access Agreement (DAA).

Project outcome

Can be published in the form research papers, patents, new therapies and other types of commonly acknowledged medical research achievements. This also includes a report on the use of samples and/or data.

Availability Information

The provider provides *Availability Information* regarding samples/data providerrequestervia BBMRI-ERIC to the requester. It is made available before the requester initiates direct inter action with the provider, in order to get access to the samples/data. '*Availability Information*' is treated as confidential by BBMRI-ERIC, i.e., it will not be disclosed to other providers.

Personal data

'Personal data' means any information relating to an identified or identifiable natural person ('*data subject'*); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (GDPR Article 4.1).

Pseudonymisation

'*Pseudonymisation*' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (GDPR Article 4.5). [Guidelines for pseudonymisation are expected to be part of the Code of Conduct on Processing of Personal Data for Purposes of Scientific Research in the Area of Health]

Human biological samples

Constituent parts of the human body, or human biological material, including organs and parts of organs, cells and tissues, and body fluids.

[1] http://bbmri-eric.eu/

[2] European Charter for Access to Research Infrastructures Principles and Guidelines for Access and Related Services, March 6. https://ec.europa.eu/research/infrastructures/pdf/2016_charterforaccessto-ris.pdf

[3] Mascalzoni D, Dove E, Rubinstein Y, Dawkins H, Kole A, Mc McCormack P, Woods S, Riess O, Schaefer F, Lochmüller H, Bartha Knoppers B, Hansson M, International Charter of Principles for Sharing Bio-specimens and Data, European Journal of Human Genetics, 2014;23:721-728.

[4] https://negotiator.bbmri-eric.eu/

I. Access Policy for CRC-Cohort

Access Policy of BBMRI-ERIC (Appendix H) applies also for access to the centrally collected data set of the CRC-Cohort, with the following exception:

Step-3 of the Access Procedure in case of CRC-Cohort is handled directly between requester and BBMRI-ERIC. For controlling access to the data set, BBMRI-ERIC has a Access Committee (see Appendix I) comprising all the contributing partner biobanks, in order to ensure the due data release approvals are in place, particularly when releasing pseudonymous (personal) data. 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 particular project. DTA also transfers liability on the requester to ensure due data security measures when processing the data.

Setup of Access Committee:

- The Access Committee is comprised of Data Manager of CRC-Cohort, one appointed person from Common Service ELSI for ethics check, and one medical expert on colorectal cancer nominated by BBMRI-ERIC.

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

- 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 project description.
- If the formal requirements are fulfilled, the medical expert assesses relevance of the project to the scope of CRC-Cohort.
- If the project is within the scope, Access Committee performs Ethics Check procedure (Appendix G) (either Expedite or Full, depending on whether sufficient previous ethics vote has been provided).
- If all the previous steps conclude successfully, Access Committee asks all the contributing partner biobanks for veto of the release. If only subset of the CRC-Cohort is requested, only those partner biobanks are contacted, who actually contributed the data specifically subject to the release. The biobanks will be given 14 days for submitting the veto (so that 1 month overall for providing the Access Committee decision is achievable).