ADOPT BBMRI-ERIC
GRANT AGREEMENT NO 676550

DELIVERABLE REPORT

Deliverable no	D2.4
Deliverable Title	Data set on samples from colorectal cancer patients
Contractual delivery month	M42 (March 2019)
Responsible Partner	BBMRI.it
Author(s)	Marialuisa Lavitrano, Petr Holub, Barbara Parodi, Outi Törnwall, Michael Hummel

DATA SET ON SAMPLES FROM COLORECTAL CANCER PATIENTS

Executive Summary

In order to create the data set on samples from colorectal cancer patients for this deliverable, a core interdisciplinary working group including biobankers, clinicians, disease registry experts, researchers, and IT experts was established. The interdisciplinary working group had the task to define the common data model. The group agreed that the data should be provided in a standardised format. Each single item of the dataset was identified based upon the criteria defined. The goal was to provide a detailed description of the data structure in order to (a) create a useful data set for colon cancer research and (b) allow for unambiguous interpretation of the data when used in the research. The group defined all the properties that are necessary for unambiguous interpretation of the results. This includes properties for each collected variable (attribute) valid not only for colorectal cancer but also for other diseases entities.

The initial model was presented to the biobanks and they were offered the possibility to review it. As a result, the data model was further refined and a final version was implemented into the software for collecting the cohort.

All the qualified EU biobanks willing to collaborate in the context of the ADOPT project were contacted to offer the opportunity to enter the pilot phase of the collection and to provide it with more detailed information regarding the process and data gathering of colorectal cancer cases; a participation letter – with information on the purpose of the collection, how it would be realised and what would be in it for the biobank (Appendix C) was sent.

The path for collection of data sets for samples from colorectal cancer patients was established in four steps: 1) Definition of the common data model, 2) Development of the CRC-Cohort Data Protection Policy, 3) Recruitment process of BBMRI-ERIC biobanks, and 4) Definition of the reimbursement model.

The collection of 10,000 datasets from multiple biobanks was one of the major tasks of ADOPT. The aim was to enable the existing, well-established biobanks in Europe to connect with BBMRI-ERIC to provide data sets and, later on, samples for future research use. The data sets were gathered, anonymized, and made centrally available for the research community to query and identify their specific research questions in colorectal cancer. The whole process of building the CRC-Cohort comprised six steps: 1) Advertisement of the effort to BBMRI-ERIC biobanks, 2) Implementation of the resulting data model in the common Metadata Repository (MDR), 3) Implementation of central data collections system (called CCDC), 4) Design and implementation of data harmonization tools, 5) Design and implementation of data quality checks, 6) Data quality improvement cycle.

COPYRIGHT NOTICE

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author*
D2.4rev1	2019-03-04	Initial version	Marialuisa Lavitrano
D2.4rev2	2019-03-18	Revised version	Petr Holub
D2.4rev3	2019-11-11	Revised version based on reviewer's comments	Marialuisa Lavitrano, Petr Holub

Glossary

BBMRI-ERIC Biobanking and BioMolecular resources Research Infrastructure - European Research Infrastructure Consortium

CCDC central data collections system

CRC cohort colorectal cancer cohort

DPP Data Protection Policy

MDR Metadata Repository

MTA/DTA Material Transfer Agreement / Data Transfer Agreement

Contents

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

Background

Dramatic advances in molecular biology have enabled rapid, comprehensive and cost-efficient analysis of clinical samples, resulting in an explosion of disease-relevant data with the potential to improve clinical care.

Fundamental discovery research is defining at the molecular level the processes that define and drive physiology and the pathogenesis of diseases. These developments, coupled with parallel advances in information technologies and electronic medical records, provide a transformative opportunity to create a new approach to diagnosis and therapy.

Research in certain diseases using genomics, proteomics, metabolomics, systems analyses, and other modern tools has begun to yield tangible medical advances, while some insightful clinical observations have spurred new hypotheses and laboratory efforts. With better integration of information both within and between research and medicine, an increasing wealth of information has been produced.

Human physiology is far more complex than any known machine. Individual humans typically differ from each other at millions of sites in their genomes. More than ten thousand of these differences are known to have the potential to alter physiology, and this estimate is certain to grow as our understanding of the genome expands. All of this new genetic information could potentially improve diagnosis and treatment of diseases by taking into account individual differences among patients.

We now have the technology to identify these genetic differences—and, in some instances, infer their consequences for disease risk and treatment response. Some successes along these lines have already occurred; however, the scale of these efforts is limited by the availability of samples and data, from patients and healthy individuals, that would be required to make possible studies to integrate molecular information with electronic medical records during the ordinary course of health care.

Realizing the full promise of precision medicine, whose goal is to provide the best available care for each individual, requires that researchers and health-care providers have access to vary large sets of samples and health and disease-related data linked to individual patients.

Well-characterised human samples and associated data are in fact unique resources for identification of new molecular features to be used as diagnostic and/or therapeutic targets. Validated accessibility of samples and data to address needs of precision medicine in colorectal cancer is one of the main goals of ADOPT. Samples (e.g., tissues [fresh, frozen or FFPE; cancer and normal tissue], serum, plasma, fcDNA, cells) and associated data of >10,000 patients from BBMRI-ERIC Members were collected.

The delivery date was extended with Amendment #2 to M42 due to the delay in collecting the colon cancer cohort. Originally it was planned in M24.

1. Methods

The planning of the work started in 2016 and it was performed in collaboration with WP3 IT-Gateway. A core interdisciplinary working group including biobankers, clinicians, disease registry experts, researchers, and IT experts was established, and teleconferences between the core working group were organized. The dataset for data associated with colorectal cancer samples was analysed within the working group to define the final data set to be collected as a part of ADOPT BBMRI-ERIC project.

The interdisciplinary working group included:

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

Table 2: Detailed composition of the interdisciplinary working group

Name	Institution	Country	Expertise
Marialuisa Lavitrano	UNIMIB	Italy	pathology, oncology, precision medicine
Michael Hummel	Charité – Universitätsmedizin	Germany	molecular diagnostic
Kurt Zatloukal	Medical University	Austria	pathology, oncology, precision medicine
Dalibor Valík	Masaryk Memorial Cancer Institute	Czech Rep.	oncology, precision medicine
Olli Carpén	University of Helsinki & Helsinki Biobank	Finland	pathology
Gerrit Meijer	Netherlands Cancer Inst.	Netherlands	oncology, precision medicine
Rudolf Nenutil		Czech Rep.	pathology
Barbara Parodi	IRCCS San Martino Hospital	Italy	quality and biobanking
Annemieke Hiemstra	Netherlands Cancer Institute	Netherlands	pathology
Mariska Bierkens	Netherlands Cancer Institute	Netherlands	molecular biology, (epi)genetic alterations
Geraldine Vink	UMC	Netherlands	oncology
Heiden Esmeralda	Charité	Germany	pre-clinical
Petr Holub	BBMRI-ERIC	Austria	medical informatics, distributed and parallel systems, big data analysis, privacy & security
Frank Ückert	DKFZ	Germany	medical informatics, data modeling, bioinformatics
Diogo Alexandre	DKFZ	Germany	medical informatics, databases,
Ondřej Vojtíšek	Masaryk University	Czech Rep.	data modeling, implementation of medical information systems

The interdisciplinary working group had the task to define the common data model. The following general data set indicated in the ADOPT BBMRI-ERIC proposal needed to be revised and made more precise:

- PERSONAL DATA (data of birth, sex, ethnicity, age at diagnosis, familial incidence)
- COMORBIDITY (Charlson index) and functional status (Karnowsky, ECOG or WHO scale)
- RISK FACTORS (direct causality: i.e., polyps; Indirect lifestyle causality: i.e., fat intake, smoking)
- FOLLOW-UP: Relapse and date of relapse
- Life status at last known contact
- DIAGNOSTIC EXAMS: Abdomino-pelvic computed tomography (CT) scan and date exams; Ultrasound; Magnetic resonance imaging (MRI); Biopsy; Colonoscopy/gastroscopy; Liver imaging; Lung imaging; Brain imaging; Skeleton imaging
- HISTOPATHOLOGICALDIAGNOSIS
- TNM staging, UICC staging
- Site of metastasis
- MOLECULAR MARKERS: Microsatellite instability, mismatch repair gene expression, KRas mutational status, EGRFR expression, other makers if applicable (e.g., APC, BRAF, p53, Ki67)
- TREATMENT: Surgery and date of surgery; Surgical radicality; Reasons for no surgery
- Pharmacotherapy: date starting (adjuvant/ neoadjuvant), end of the treatment;
- Reasons for no chemotherapy; Targeted Treatment: date starting and Type
- of targeted treatment
- TREATMENT RESPONSE
- Overall functional status for the participant/general condition including pain status
- FOLLOW-UP and SURVIVAL.

The group agreed that the data should be provided in a standardised format. Each single item of the dataset was identified based upon the criteria defined. The goal was to provide a detailed description of the data structure in order to (a) create a useful data set for the colon cancer research, (b) allow for unambiguous interpretation of the data when used in the research.

The group defined all the properties that are necessary for unambiguous interpretation of the results. This includes the following properties for each collected variable (attribute) valid non only for colorectal cancer but also for other diseases entities (see D2.5):

- Unique label of the variable
- Short description (label) of the variable - to be used in forms
- Semantics = definition of meaning
  - This includes references to existing clearly defined official standards or community "standards", including existing ontologies
  - We will use this for ontologizing the data model, in order to make it "machine readable" (allowing for correct interpretation of the data in automated processing workflows)
- Syntax
  - including data type (elementary types such as boolean, float, integer, free text, specifically structured text, etc., array or lists of elementary types)
  - including coding (e.g., IEEE 754 for floats, regular expressions for structured text)
- List of allowed units
  - including their conversion algorithms (with "non-existent" and "unknown" interim options)

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

- Level: REQUIRED, OPTIONAL, RECOMMENDED
  - REQUIRED means the data can't be entered at all without this item being provided
  - OPTIONAL means data may or may not be provided, but the item will be ready for inputting the data in as part of the data model,
  - RECOMMENDED is a special subclass of the OPTIONAL, which is highly recommended to be filled in (intended for items where we need the data but where we know that some sources won't be able to fill this in and we still want such data not being discarded as invalid)
- Relation to entities (patient, examination, etc.) - will be used for developing the formal model.

The development of the data model was done in five steps:

1. Basic consensus on collected attributes among the medical experts.
   Deadline: April 30, 2016

2. Development of formal model including entities, their attributes and their mutual relations by IT experts.

3. Review of the formal model by the joint group of medical and IT experts.

4. Approval of the resulting formal model by the BBMRI-ERIC Management Committee (used also as project management board in ADOPT BBMRI-ERIC project).
   Deadline: June 30, 2016

5. Development of the data collection application (ADOPT WP3) for manual data collection (manual data collection itself will be done within ADOPT WP2).
   Deadline: September 30, 2016.

The initial model was presented to the biobanks and they were offered the possibility to review it. This resulted in a review document that was maintained as shown in Appendix A and the data model was further refined. The entire process lasted until January 2018; from then onward, the model was immutable and its final version was implemented into the software for collecting the cohort. Based on the final model, instructions for biobanks were prepared (Appendix B), detailing also how to construct the XML data structure for import into the central DB.

All the 60 qualified EU biobanks included in the list (see D2.2) of those willing to collaborate in the context of the ADOPT project were re-contacted to offer the opportunity to enter the pilot phase of the collection. In order to provide them with more detailed information of the process and data gathering of colorectal cancer cases, a participation letter – with information of the purpose of the collection, how it would be realised and what would be in it for the biobank (Appendix C) was sent. Information about the timelines and next steps needed in order to enter the pilot phase of the collection and data items that were to be collected from each existing colorectal cancer case were gathered. Direct connection with the relevant people in the biobanks (manager/IT personnel/researcher/MD expert) was established.

A second wave of mapping the biobanks for the colorectal cancer collection was also performed and an updated list of qualified EU biobanks was created (Appendix D).

The path for collection of data sets for samples from colorectal cancer patients was established in four steps:

1. **Definition of the common data model** created by the interdisciplinary working group of the medical and lT experts. The data model focused on unambiguous definition of the data structure so that it could be implemented in lT systems, and on defining which parts of the data model are required to obtain data set meaningful for medical research.
2. **Development of CRC-Cohort Data Protection Policy** in order to allow the contributing biobanks to request approval by their governing bodies. This policy was developed by Petr Holub, lrene Schlünder, Kurt Zatloukal, Outi Tornwall, Marialuisa Lavitrano, and Michael Hummel. lt was developed between 2017-06-19 and 2017-10-16, when the final version 1.1 was released after discussion and approval by BBMRI-ERIC Management Committee. This version was used in the data collection process. The full version of the policy is in D2.4 Appendix III. The proposed policy was consulted with the biobanks that indicated their interest in participating in the CRC-Cohort. The CRC-Cohort Data Protection Policy was published as ADOPT D2.3 Appendix III.
3. **Recruitment process of BBMRI-ERIC biobanks** and obtaining feedback on the proposed data model, given the heterogeneity of European health care and medical standards. The evaluation of the state-of-the-art of automated extraction of clinical data in years 2015-2016 resulted in a decision to use a backup plan already prepared in the ADOPT project proposal, to focus on semi-automated extraction for those biobanks with sufficiently advanced lT and to support all biobanks in automated transformations of already structured data. A survey among the biobanks revealed that most of the biobanks already had substantial parts of the data available in structured form, with the typical notable exception of treatment and responses to treatments; hence retrieval of the data for CRC-Cohort was almost always a mixture of machine processing of structured data and manual completion of missing data and fixes of data quality. Hence almost all biobanks qualified into semi-automated extraction mode.
4. **Definition of the reimbursement model.** The resources available for reimbursing manual work of the biobank were combined with the lT budget for automated processing of records. Based on the decision of the Management Committee and Director General, there was also a bonus proposed for early contributors to the CRC-Cohort.

The collection of 10,000 datasets from multiple biobanks was one of the major tasks of ADOPT. The aim was to enable the existing, well-established biobanks in Europe to connect with BBMRI-ERIC to provide data sets and, later on, samples for future research use. The data sets were gathered, anonymized, and made available centrally for the research community to query and identify their specific research questions in colorectal cancer.

Preparation for this massive effort was accelerated after the data model was created and the mapping of qualified biobanks in Europe was completed. The preparation work consisted of drafting and sending two different letters to the biobanks with detailed information of process and writing a Data Protection Policy (DPP) while simultaneously maintaining the open discussion line with the biobanks. In March 2017, the 1st letter (Appendix C) was sent to 75 different biobanks in 17 countries: Austria, Belgium, Czech Republic, Cyprus, Estonia, Finland, France, Germany, Italy, Malta, the Netherlands, Norway, Sweden, Switzerland, Poland, Turkey and UK. The purpose was to inform the biobanks of the next steps, to collect the information on the availability and format of the data and therefore also to ensure the biobanks' abilities to fulfil the inclusion criteria.

Based on the availability of the data and the feedback from the biobanks, the biobanks were divided into three groups according to their method of participation: manual collection of data sets, semi-automated collection of datasets and automated data collection. For each group, the 2nd letter was prepared and sent in July 2017 informing them of the next steps together with a Data Protection Policy document, which was a major collaborative effort driven by WP2 and WP3. It gathered the objectives of the CRC cohort, its legal framework and basic organizational aspects. It described the data collection and integration process, together with measures for quality checking and assurance. Access modes for the data set were also discussed and an overview of tools on which the implementation of the CRC-Cohort relies was provided.

The whole process of building the CRC-Cohort comprised six steps:

1. **Advertisement of the effort to BBMRI-ERIC biobanks** and preliminary inquiry of their interest in participating.
2. **Implementation of the resulting data model in the common Metadata Repository (MDR)**, where it is in machine-readable form to be used by applications. The data model defined by the interdisciplinary expert working group was implemented in the MDR where it is available via an APl for access by other components of the CRC-Cohort ecosystem of lT tools. The MDR instance hosting the data model is publicly available at https:IImdr.osse-register.deIview.xhtml?namespace=ccdg. Availability of the data model in a machine-readable structure is a prerequisite in order to have the whole system FAlR compliant in the future.
3. **Implementation of central data collections system (called CCDC),** including database, web-based user interface for manual contributions and APl to allow programmatic imports of the data. The central CCDC system for collecting data was implemented based on open-source software coming from OSSE Project2. The software was extended adequately to support the CRC-Cohort and to feature both graphical user interface {via web) or an APl for programmatic upload of the data into the system (see D2.7).
4. **Design and implementation of data harmonization tools** was designed to support the conversion process from common tabular files (Excel, CSV/TSV files, etc.).
5. **Design and implementation of data quality checks** in collaboration between expert pathologists and lT experts. Statistical inspection of initially collected data showed need for providing complex data quality reporting tool that would help contributing biobanks to detect problems in the delivered data and handle those issues. A system of data quality checks was developed as a central service running on the central database (see D2.7).
6. **Data quality improvement cycle**, where results of the checks were fed back to the contributing biobanks, and the centrally collected data set was updated based on updated data from the biobanks.

The effort took place since March 2016, starting with the definition of the data set to be collected, with the main data collection period running from January 2018 until March 2019.

The effort turned out to be organizationally very demanding, far exceeding the original expectations, for a number of reasons: there are huge differences in availability of structured in-depth data in different European countries {also stemming from ability of biobanks to connect to national registries collecting these data), differences in organizational requirements, as well as differences in availability of lT expertise to manipulate the data at source.

2. Results

2.1. 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 (C18.1 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 MTA/DTA.

2.2. Overview of the data model

The entity-relation diagram of the resulting data model is shown in Figures 1 and 2. Overview of the variables is shown in Figures 3 to 5.

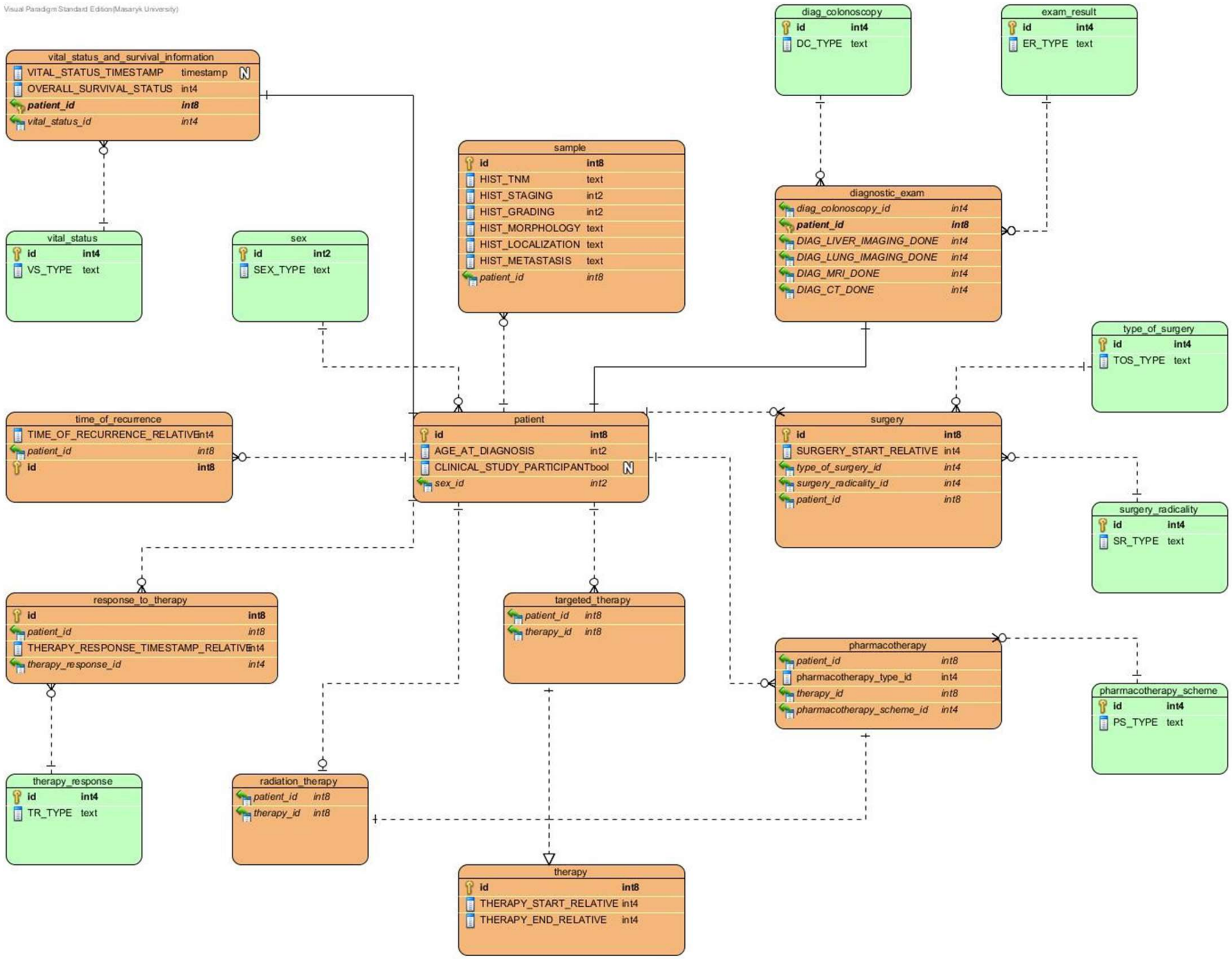


Figure 1: Patient-related entity-relation diagram of the data model

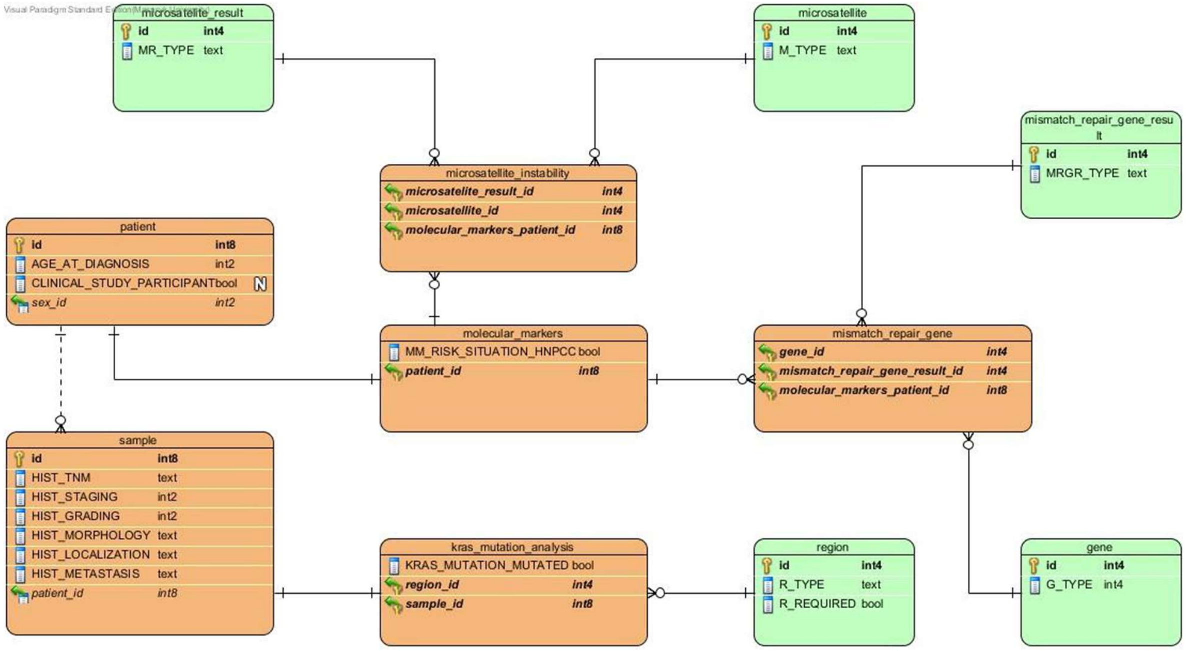


Figure 2: Entity-relation diagram of the data model related to the molecular markers for a patient

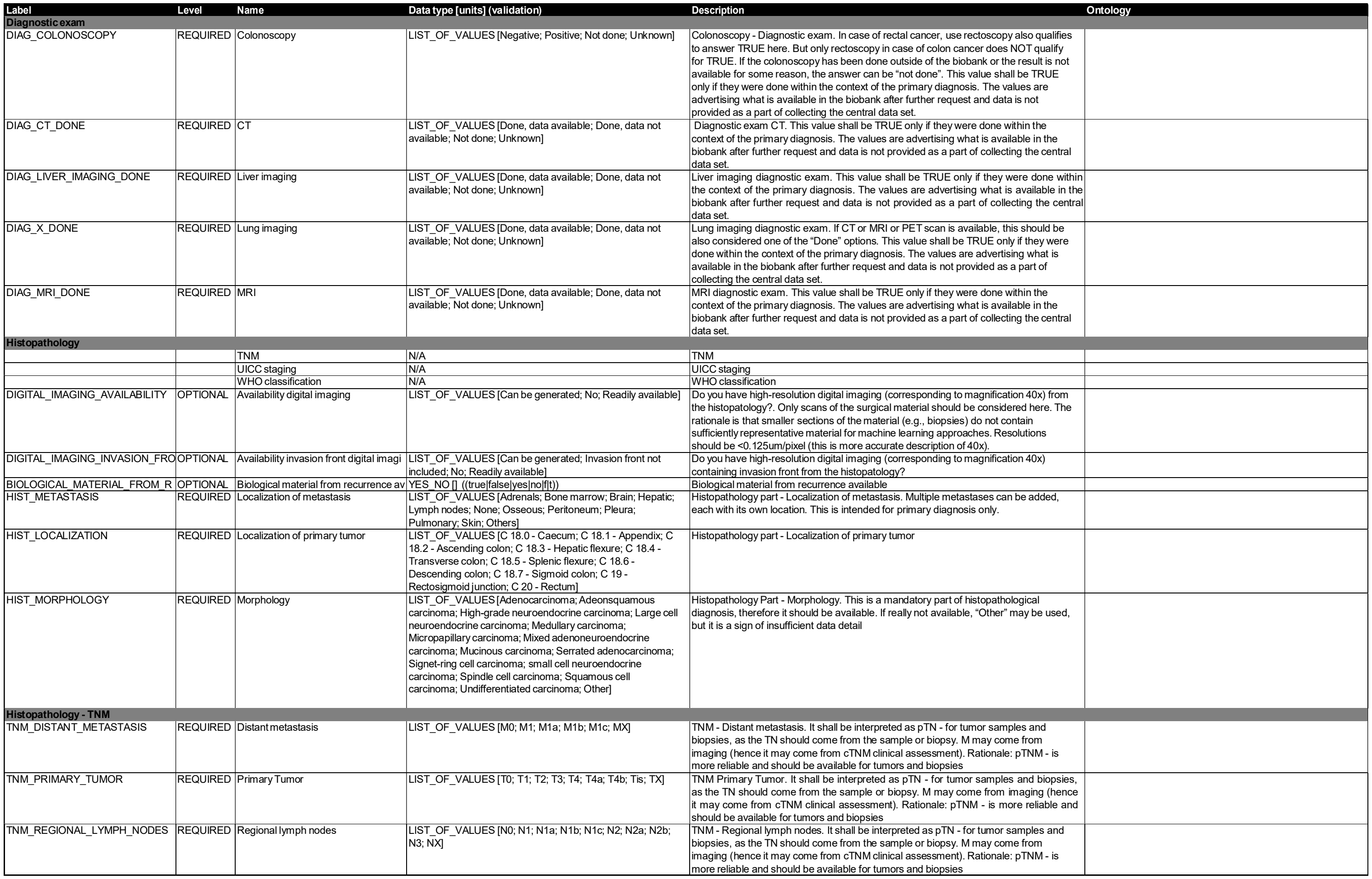
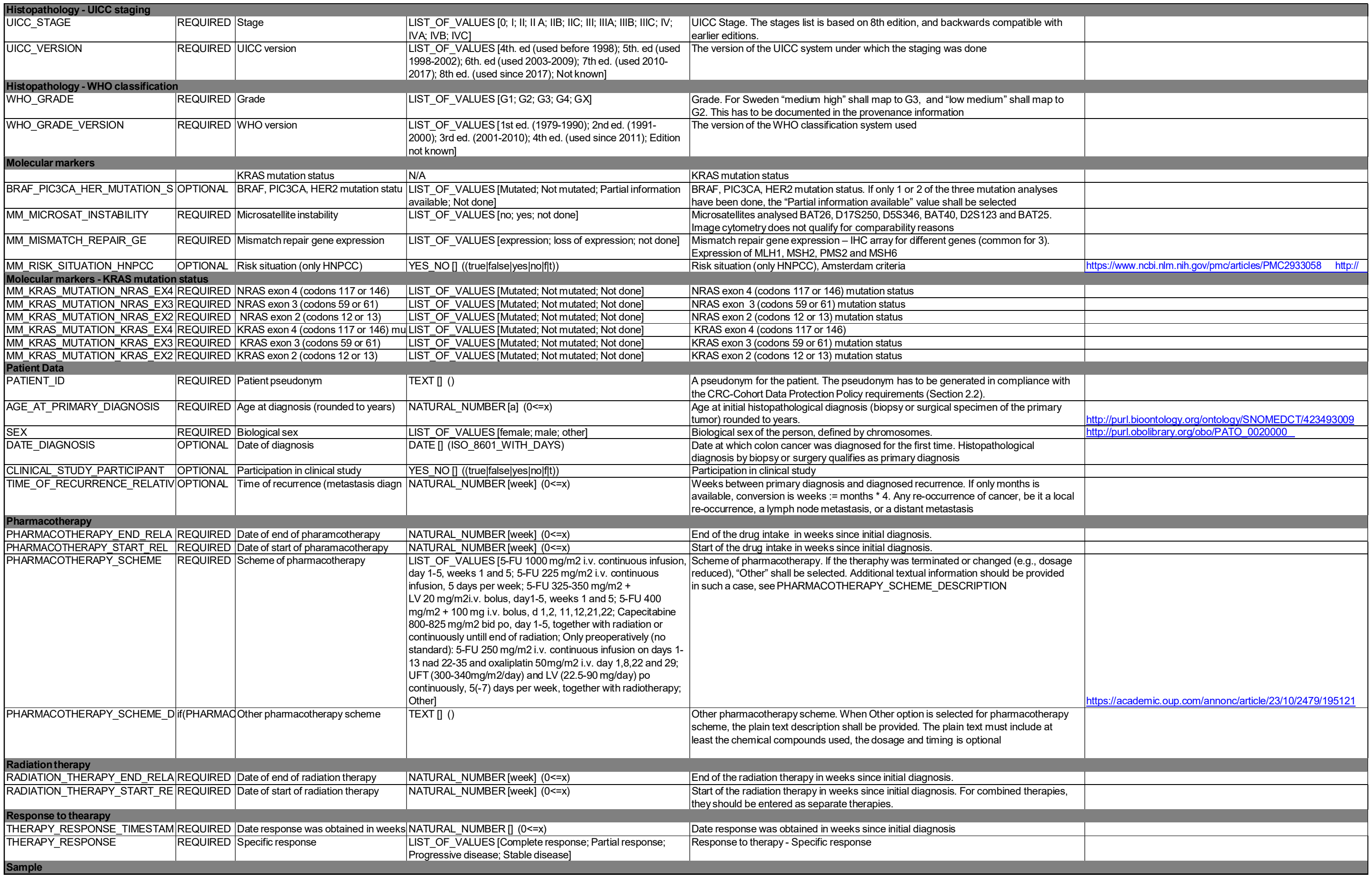


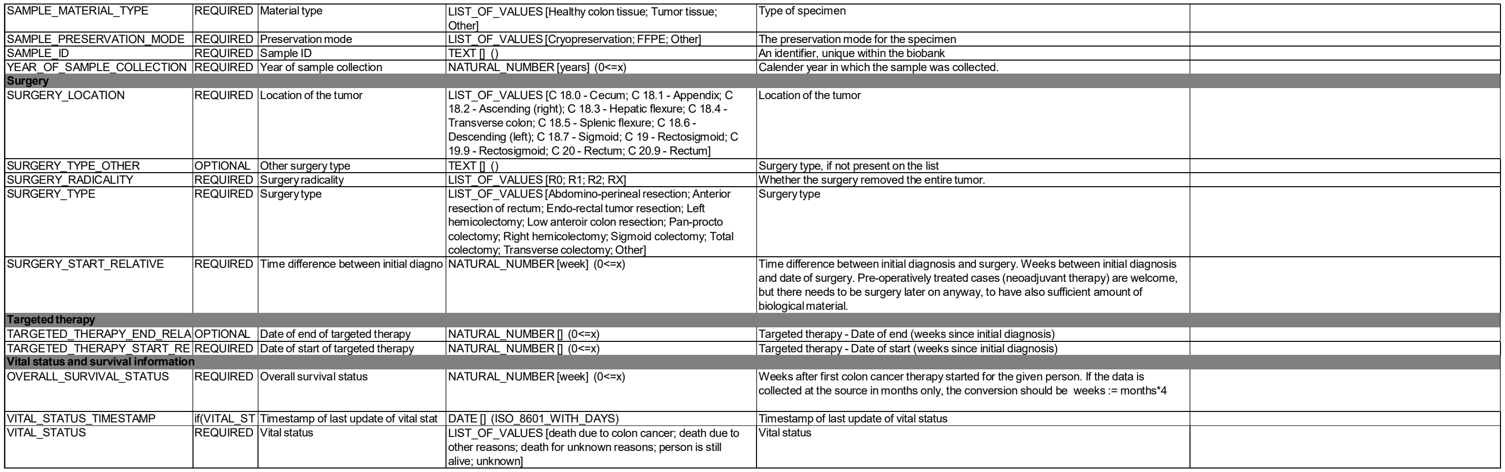
Figure 3: Overview of defined variables in the data model – part I

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_R	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_RELA	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					

Figure 4: Overview of defined variables in the data model – part II



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 diagno	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 stat	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	

Figure 5: Overview of defined variables in the data model – part III

2.1. Details on variables

This is the structure of variables that came out of the expert WG before feedback from the biobanks. Note that some clarifications were subsequently introduced based on the feedback from the biobanks; the resulting refined model is as shown above in Figures 3 to 5.

- Sex:
  - Label: `SEX`
  - Short description: Biological sex
  - Semantics:
    * Biological sex of the person, defined by chromosomes.
    * `http://purl.obolibrary.org/obo/PATO_0020000`
  - Syntax: male, female (only 2 values allowed)
  - Units: n/a
  - Level: REQUIRED
- Participation in clinical study
  - 1..1 to patient
  - Label: `CLINICAL_STUDY_PARTICIPANT`
  - Semantics:
    * Participant of clinical study.
  - Syntax: boolean
  - Units: n/a
  - Level: RECOMMENDED
- Age at primary diagnosis:
  - Label: `AGE_AT_PRIMARY_DIAGNOSIS`
  - Short description: Age at diagnosis (rounded to years)
  - Semantics:
    * Age at initial histopathological diagnosis (biopsy or surgical specimen of the primary tumor) rounded to years.
    * `http://purl.bioontology.org/ontology/SNOMEDCT/423493009`
  - Syntax: integer
  - Units: years since birth
  - Level: REQUIRED
- Time of recurrence (metastasis):
  - 0..n related to the patient
  - Label: `TIME_OF_RECURRENCE_RELATIVE`
  - Short description: Time of recurrence (metastasis diagnosis)
  - Semantics:
    * Weeks between primary diagnosis (`AGE_AT_PRIMARY_DIAGNOSIS`) and diagnosed recurrence

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

    - ∗ If only months is available, conversion is weeks := months * 4
  - Syntax: integer
  - Units: weeks since primary diagnosis
  - Level: OPTIONAL

- Vital status and survival information
  - 1..1 to person (= REQUIRED)
  - Vital status
  - Label: `VITAL_STATUS`
  - Semantics: living or deceased
  - Syntax:
    - ∗ list (`ALIVE`= … person is still alive, `DEATH_COLON_CANCER` = death due to colon cancer, `DEATH_OTHER` = death due to other reasons, `DEATH_UNKNOWN_REASON` = death for unknown reasons, `UNKNOWN` = unknown)
  - Units: n/a
  - Level: REQUIRED

- Timestamp of last update of vital status
  - Label: `VITAL_STATUS_TIMESTAMP`
  - Semantics:
  - Timestamp of last update of vital status
  - Syntax: timestamp compliant to ISO 8601
  - Units: n/a
  - Level: REQUIRED if `VITAL_STATUS` != `UNKNOWN`

- Overall survival status
  - Label: `OVERALL_SURVIVAL_STATUS`
  - Semantics:
    - ∗ 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
  - Syntax: integer
  - Units: weeks
  - Level: REQUIRED

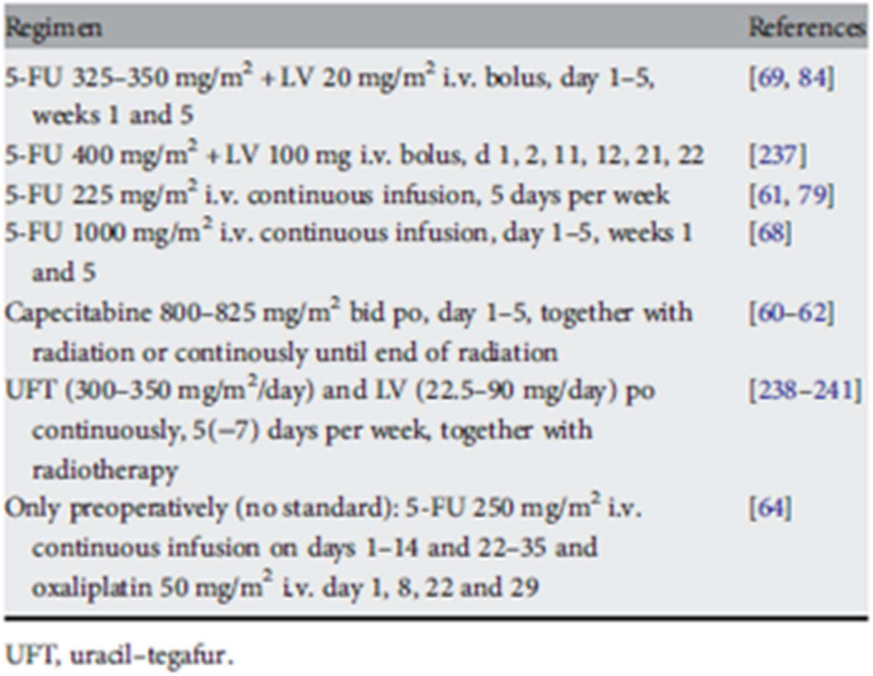
- Surgery: aggregate object
  - 0..n - patient to surgery
  - Time difference between initial diagnosis and surgery:
  - Label: `SURGERY_START_RELATIVE`
  - Semantics:
    - ∗ Weeks between initial diagnosis and date of surgery.
  - Syntax: integer
  - Units: weeks
  - REQUIRED

- Surgery radicality:

  - Label: SURGERY_RADICALITY
  - Semantics:
  - Whether the surgery removed the entire tumor.
  - Syntax: list (RX, R0, R1, R2)
  - Units: n/a
  - Level: REQUIRED
  - Type of surgery:
  - Label: SURGERY_TYPE
  - Semantics: “OTHER” value may allow for optional “please specify” free text option
  - Syntax: list (RIGHT_HEMICOLECTOMY, LEFT_HEMICOLECTOMY, TRANSVERSE_COLECTOMY, SIGMOID_COLECTOMY, TOTAL_COLECTOMY, PAN-PROCTO_COLECTOMY, LOW_ANTERIOR_COLON_RESECTION, ANTERIOR_RESECTION_OF_RECTUM, ABDOMINO-PERINEAL_RESECTION, ENDO-RECTAL_TUMOR_RESECTION, OTHER)
  - Units: n/a
  - Level: REQUIRED

- Pharmacotherapy:

  - 0..n to patient
  - REQUIRED if occurred
  - Date of start:
    - Label: PHARMACOTHERAPY_START_RELATIVE
    - Semantics: start of the drug intake in weeks since initial diagnosis.
    - Syntax: integer
    - Units: weeks
    - Level: REQUIRED
  - Date of end:
    - Label: PHARMACOTHERAPY_END_RELATIVE
    - Semantics: end of the drug intake in weeks since initial diagnosis.
    - Syntax: integer
    - Units: weeks
    - Level: REQUIRED
  - Scheme of pharmacotherapy:
    - Label: PHARMACOTHERAPY_SCHEME
    - Semantics:
    - Pointer to one of the rows of the following table [2, Table 11]:
    - Syntax: list (rows from the table above)
    - Units: N/A
    - Level: REQUIRED



Regimen	References
5-FU 325–350 mg/m^2 + LV 20 mg/m^2 i.v. bolus, day 1–5, weeks 1 and 5	[69, 84]
5-FU 400 mg/m^2 + LV 100 mg i.v. bolus, d 1, 2, 11, 12, 21, 22	[237]
5-FU 225 mg/m^2 i.v. continuous infusion, 5 days per week	[61, 79]
5-FU 1000 mg/m^2 i.v. continuous infusion, day 1–5, weeks 1 and 5	[68]
Capecitabine 800–825 mg/m^2 bid po, day 1–5, together with radiation or continously until end of radiation	[60–62]
UFT (300–350 mg/m^2/day) and LV (22.5–90 mg/day) po continuously, 5(–7) days per week, together with radiotherapy	[238–241]
Only preoperatively (no standard): 5-FU 250 mg/m^2 i.v. continuous infusion on days 1–14 and 22–35 and oxaliplatin 50 mg/m^2 i.v. day 1, 8, 22 and 29	[64]

UFT, uracil-tegafur.

- Targeted therapy:
  - 0..n to patient
    - REQUIRED if occurred
  - Date of start:
    - Label: `TARGETED_THERAPY_START_RELATIVE`
    - Semantics: start of the drug intake in weeks since initial diagnosis.
    - Syntax: integer
    - Units: weeks
    - Level: REQUIRED
  - Date of end:
    - Label: `TARGETED_THERAPY_END_RELATIVE`
    - Semantics:
    - end of the drug intake in weeks since initial diagnosis.
    - Syntax: integer
    - Units: n/a
    - Level: OPTIONAL
- Radiation therapy:
  - 0..1 to patient
    - REQUIRED if occurred
  - Date of start:
    - Label: `RADIATION_THERAPY_START_RELATIVE`
    - Semantics:

        - start of the radiation therapy in weeks since initial diagnosis.
        - Syntax: integer
        - Units: weeks
        - Level: REQUIRED
    - Date of end:
        - Label: `RADIATION_THERAPY_END_RELATIVE`
        - Semantics:
        - end of the radiation therapy in weeks since initial diagnosis.
        - Syntax: integer
        - Units: weeks
        - Level: REQUIRED

- Response to therapy
    - 0..n to patient
    - 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
    - Label: `THERAPY_RESPONSE`
    - Semantics: Therapy response according to RECIST criteria [1].
    - Syntax: list (`PROGRESSIVE_DISEASE`, `STABLE_DISEASE`, `PARTIAL_RESPONSE`, `COMPLETE_RESPONSE`)
    - Units: n/a
    - Level: REQUIRED (only if the response exists - see overall 0..n relation to patient)

- Specific response timestamp
    - Label: `THERAPY_RESPONSE_TIMESTAMP_RELATIVE`
    - Semantics:
    - Timestamp when the therapy response was obtained, in weeks relative to the initial diagnosis
    - Weeks := months * 4 (if only months are available)
    - Syntax: integer
    - Units: weeks
    - Level: REQUIRED (only if the response exists - see overall 0..n relation to patient)

- Molecular markers
    - Microsatellite instability
        - 1..1 to person
        - Label: `MM_MICROSAT_INSTABILITY`
        - Semantics:
        - Microsatellites analysed BAT26, D17S250, D5S346, BAT40, D2S123 and BAT25
        - Syntax: list (`NOT_DONE`, `NO`, `YES`) – SINGLE-VALUE
        - Units: N/A
        - Level: REQUIRED
    - Mismatch repair gene expression – IHC array for different genes (common for 3)

    - 1..1 to person
    - Label: `MM_MISMATCH_REPAIR_GE`
    - Semantics:
    - existing guidelines - immunohistochemistry
    - Expression of MLH1, MSH2, PMS2 and MSH6
    - Syntax: list (`NOT_DONE`, `EXPRESSION`, `LOSS_OF_EXPRESSION`) – SINGLE-VALUE
    - Units: N/A
    - Level: REQUIRED
  - Risk situation (only HNPCC)
    - 1..1 to person
    - Label: `MM_RISK_SITUATION_HNPCC`
    - Semantics:
    - Amsterdam criteria (Vasen HF, Watson P, Mecklin JP, Lynch HT (1999). ”New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC”. Gastroenterology 116 (6): 1453–6) OR Bethesda criteria (Umar A, Boland CR, Terdiman JP, et al. (2004). ”Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability.”. J Natl Cancer Inst. 96 (4): 261–268)
    - Syntax: boolean (`TRUE` = risk situation)
    - Units: N/A
    - Level: OPTIONAL
  - KRAS mutation status
    - 1..1 to sample
    - Label: `MM_KRAS_MUTATION`
    - Semantics:
    - Syntax: list (`NOT_DONE`, list of defined exons) – MULTI-VALUE
    - KRAS ex2 (codons 12 or 13), ex3 (codon 59 or 61), ex4 (codon 117 or 146); NRAS ex2 (codon 12 or 13), ex3 (codon 59 or 61), ex4 (codon 117 or 146)
    - For each of those, we need to know boolean yes/no (mutated/non-mutated); insertions/deletions/indels do not need to be considered
    - Units: N/A
    - Level: REQUIRED
  - BRAF, PIC3CA. HER2 mutation status
    - 1..1 to sample
    - Semantics:
    - Syntax: list of lists – one list of values per mutation (`NOT_DONE`, `MUTATED`, `NON_MUTATED`) - default is `NOT_DONE` since people will mostly not have these done
    - List of relevant genes: BRAF, PIC3CA. HER2
    - Units: N/A
    - Level: OPTIONAL

- Histopathology part

  - 1..1 - to sample
  - TNM
    - Label: `HIST_TNM`
    - Semantics:
    - Syntax: UICC standard
    - Units: N/A

    * Level: REQUIRED
- UICC staging
    * Label: `HIST_STAGING`
    * Semantics:
    * standard – but changes over the time
    * Syntax: UICC (stage I to IV)
    * Units: n/a
    * Level: REQUIRED
    * Note: must include definition of the version of UICC standard used (can be implemented indirectly by providing date of determination of value)
- WHO grading
    * Label: `HIST_GRADING`
    * Semantics:
    * standard – but changes over the time
    * Syntax: WHO (G1 to G4)
    * Units: n/a
    * Level: REQUIRED
    * Note: must include definition of the version of UICC standard used (can be implemented indirectly by providing date of determination of value)
- Morphology
    * Label: `HIST_MORPHOLOGY`
    * Semantics:
    * Syntax: UICC standard
    * Units: n/a
    * Level: REQUIRED
- Localization
    * Label: `HIST_LOCALIZATION`
    * Semantics:
    * standard – ICD-10
    * Syntax: ICD-10 standard – C18.1 to C18.7, C19, C20
    * Units: n/a
    * Level: REQUIRED
- Metastasis
    * Label: `HIST_METASTASIS`
    * Semantics:
    * localization of metastasis
    * Based on ICD-10
    * Syntax: UICC standard
    * Units: n/a
    * Level: REQUIRED
    * Note: yes/no is part of TNM

- Diagnostic exam (1..1 relation to patient)

  – Colonoscopy
    ∗ Label: `DIAG_COLONOSCOPY`
    ∗ Semantics:
    ∗ whether colonoscopy was done
    ∗ Syntax: list (`NOT_DONE`, `POSITIVE`, `NEGATIVE`)
    ∗ Units: n/a
    ∗ Level: REQUIRED
    ∗ Note: possible relation/interaction with the localization in histopathological diagnosis
    ∗ Array of diagnostic methods (liver imaging, lung imaging, MRI, CT)
  – Label: `DIAG_X_DONE` – X ∈ {`LIVER_IMAGING`, `LUNG_IMAGING`, `MRI`, `CT`}
    ∗ Semantics:
    ∗ whether given diagnostics was done
    ∗ Syntax: list (`NOT_DONE`, `DONE_DATA_AVAILABLE`, `DONE_DATA_NOT_AVAILABLE`, `UNKNOWN`)
    ∗ Units: n/a
    ∗ Level: REQUIRED

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

3. Discussion and Conclusions

The rise of data-intensive biology, advances in information technology, and changes in the way health care is delivered have created a compelling opportunity to improve the diagnosis and treatment of disease.

Biology has acquired the capacity to systematically compile molecular data on a scale that was unimaginable 20 years ago. Diverse technological advances make it possible to gather, integrate, analyse, and disseminate health-related biological data in ways that could greatly advance both biomedical research and clinical care. Meanwhile, the magnitude of the challenges posed by the sheer scientific complexity of the molecular influences on health and disease are becoming apparent and suggest the need for powerful research resources. All these changes provide an opportunity for the biomedical science and clinical communities to come together to improve both the discovery of new knowledge and health-care delivery.

Studies that aid in the understanding of cancer on a molecular level have provided important tools for genetic testing for high-risk familial forms of the disease, predictive markers for selecting patients for certain classes of drug therapies, and molecular diagnostics for the non-invasive detection of early cancers. In addition, biologic pathways that could form the basis of new therapeutic agents have been identified in different cancers. Although some high-frequency mutations are attractive targets for drug development, common signalling pathways downstream from these mutations may also be tractable as therapeutic targets.

Recent progress in molecular assays for the early detection of cancers indicates that understanding the genes and pathways that control the earliest steps of the disease and individual susceptibility can contribute to clinical management in the near term. An understanding of the signals that dictate the metastatic phenotype will provide the information necessary to develop drugs to control or prevent advanced disease. The considerable recent advances encourage both clinicians and researchers to believe that improvements in our knowledge of the molecular basis of cancers will continue to reduce the burden of this disease.

The ambition of BBMRI-ERIC is to implement a world-leading Research Infrastructure for biomedical research in Europe – a true gateway for health. Well-characterised human samples and associated data are unique resources for identification of new molecular features to be used as diagnostic and/or therapeutic targets.

The collection of >10,000 datasets from multiple biobanks is one of the major tasks of ADOPT. The aim is to enable the existing, well-established biobanks in Europe to connect with BBMRI-ERIC to provide data sets and, later on, samples for future research use.

We defined the data set to be collected and provided a detailed description of the data structure in order to (a) create a useful dataset for the colon cancer research, (b) allow for unambiguous interpretation of the data when used in the research.

This effort of forming a cohort with existing 10,480 colorectal cancer cases with detailed pathological and clinical data and available tissue samples demonstrated the feasibility of large-scale collaboration within BBMRI-ERIC and generated a yet unprecedented resource for medical research.

The CRC-Cohort will become a permanent asset of the BBMRI-ERIC research infrastructure after the end of the project in order to enable research to improve treatment of colorectal cancer.

The data collection will provide broad European coverage and sufficient number of research participants in order to enable research that had previously been impossible. The cohort should enable a large spectrum of different types of research and is, therefore, not designed for or restricted to a specific research question.

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 histo-pathological sections together with outcome data for the development of so-called imaging biomarkers by machine learning.
- to establish a benchmark data set for evaluation of quality of anonymization techniques and related residual risk of re-identification by BBMRI-ERIC.
- 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 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.

The main lesson learned from ADOPT was that collecting cohorts like this requires that biobanks:

- are connected to the sources of clinical data
- are connected to the relevant expertise (pathological, clinical, etc.)
- have sufficient time to deal with identified quality issues
- are willing to iteratively improve data quality
- have substantial IT resources (this is true for BBMRI-ERIC headquarters as well)
- have capacity and expertise to improve data quality

In addition, most of the implemented statistic checks requires availability of substantial amounts of data to provide meaningful results. It is difficult to extract structured data from originally unstructured clinical data. Moreover, complex legal and procedural issues slow down the process; resources must be dedicated for these purposes as well.

Bibliography

[1] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, et al. "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)". In: *European journal of cancer* 45.2 (2009), pp. 228–247. URL: `https://ctep.cancer.gov/protocoldevelopment/docs/recist_guideline.pdf`.

[2] K. Jordan, H. J. Schmoll, C. J. van de Velde, J. Balmana, J. Regula, I. D. Nagtegaal, R. G. Beets-Tan, F. Ciardiello, P. Hoff, D. Kerr, C. Köhne, E. Van Cutsem, R. Labianca, T. Price, W. Scheithauer, A. Sobrero, J. Tabernero, D. Aderka, S. Barroso, G. Bodoky, J. Y. Douillard, H. El Ghazaly, A. Stein, D. Arnold, J. Gallardo, A. Garin, A. Meshcheryakov, R. Glynne-Jones, D. Papamichail, P. Pfeiffer, I. Souglakos, S. Turhal, A. Cervantes, V. Valentini, B. Glimelius, K. Haustermans, and B. Nordlinger. "ESMO Consensus Guidelines for management of patients with colon and rectal cancer. A personalized approach to clinical decision making". In: *Annals of Oncology* 23.10 (Oct. 2012), pp. 2479–2516. ISSN: 0923-7534. OOI: `10.1093/annonc/mds236`. eprint: `http://oup.prod.sis.lan/annonc/article-pdf/23/10/2479/488404/mds236.pdf`. URL: `https://dx.doi.org/10.1093/annonc/mds236`.

A. Data Model Review Log

This appendix provides a detailed log of comments obtained in the data model review by the biobanks, done between January 2017 and January 2018.

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

CRC cohort - medical and data Qs

#	Data Item	List of values	Question	Resolution	Action	Answer to the biobank
1	DIAG_COLONOSCOPY		Does this include colonoscopy only, or can rectoscopy also be documented (for example, in rectal carcinoma)?	[COMMENT CHANGE] 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.	Add the comment to the MDR. [DONE] Answer to Frankfurt (Gabriele Husman) [DONE]	For rectal cancer the retroscopy qualifies, but for other type of cancers not.
2	HIST_METASTASIS	LIST_OF_VALUES [Localization of metastasis – Others, Localization of metastasis – Skin, Localization of metastasis – Adrenals, Localization of metastasis – Peritoneum, Localization of metastasis – Pleura, Localization of metastasis – Bone marrow, Localization of metastasis – Lymph nodes, Localization of metastasis – Brain, Localization of metastasis – Hepatic,	1) If there are no metastases, can the field remain empty or must it be filled in any case (Localization of meta stasis – None)? 2) Is it meaningful to collect so multiple values?	[COMMENT CHANGE] Multiple metastasis values can be added (= supported by the data model). [IGNORE] The only medically critical value is single distant liver metastasis. [IGNORE] Possibly we may add cTNM for each metastasis as OPTIONAL parameter, see discussion of TNM below. [IGNORE] Suggestion from Sebastian to remove "Localization of metastasis -" prefix from the values should be ignored; that is an artifact from the MDR that enables to have multiple values per	Add a description of entities and relations and cardinalities into the data model (separate page as a diagram - and possibly also into notes on categories of attributes). [DONE] Answer to Frankfurt (Gabriele Husman) [DONE]	1) The value: "Localization of metastasis – None" should be used in this case 2) Yes, because multiple metastasis values can be added.
			3)			3)

		Localization of metastasis − Osseous, Localization of metastasis − Pulmonary, Localization of metastasis − None]		attribute and won't display in the CCDC UI. The prefixes shall be stripped as a part of the XML -> XLS processing of the export from MDR.		
3	HIST_LOCALIZATION	LIST_OF_VALUES [Localization of primary tumor - C 20, Localization of primary tumor - C 19, Localization of primary tumor - C 18.7, Localization of primary tumor - C 18.6, Localization of primary tumor - C 18.5, Localization of primary tumor - C 18.4, Localization of primary tumor - C 18.3, Localization of primary tumor - C 18.2, Localization of primary tumor - C 18.1]	Unfortunately BBMRI database does not allow the documentation of **C18.0** = caecum. Is it possible to extend the database? Otherwise, cases will get lost because it is not possible to document the tumor localization C18.0.	[INCLUSION CRITERIA CHANGE] Include C18.0. [VALUE CHANGE] Add C18.0.	[Done.] Answer to Frankfurt (Gabriele Husman) [DONE]	Yes, the database will be expanded.
4	TNM_DISTANT_METASTASIS TNM_REGIONAL_LYMPH_NODES TNM_PRIMARY_TUMOR		In TNM the prefix "c" or "p" is missing. A distinction can not be made between a clinical TNM in neoadjuvant therapy and a pathological TNM after OP.	[COMMENT CHANGE] Attribute values shall be interpreted as pTN - for tumor samples and biopsies: the TN values shall come from the sample or biopsy. M may come from imaging (hence it may come	Add clarification into the MDR. [DONE] Answer to Frankfurt (Gabriele Husman)	Attribute values shall be interpreted as pTN - for tumor samples and biopsies: the TN values shall come from the sample or biopsy. M may come from imaging (hence it may come from cTNM clinical

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

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

				from cTNM clinical assessment). Rationale: pTNM - is more reliable and should be available for tumors and biopsies. [IGNORE] cTNM - would have to be elsewhere in the model, it can be used for metastases, but we don't collect it as of now.	[DONE]	assessment). Rationale: pTNM - is more reliable
5	UICC_STAGE	LIST_OF_VALUES [Stage - IVB, Stage - IVA, Stage - IIIC, Stage - IIIB, Stage - IIIA , Stage - IIB, Stage - II A, Stage - I, Stage - 0]	According to TNM 7th edition, stadium IIC and III are missing.	[VALUE ADDITION] Add missing values: IIC, IVC, II, III, IV. [Highlight the edition which we are using. - That's what we already have in the UICC_VERSION attribute.] [VALUE ADDITION for TNM_PRIMARY_TUMOR] Add T4 (without any letter). [VALUE ADDITION for NM_DISTANT_METASTASIS] Add M1c.	Add 2c for current UICC to MDR For UICC 1997, add 1, 2, 3, 4. [DONE] Answer to Frankfurt (Gabriele Husman) [DONE]	We will add the missing values and highlight the edition used for UICC staging
6	UICC_VERSION	LIST_OF_VALUES [4th edition or earlier, 5th edition, 6th edition, 7th edition, Not known]	Please also add the current 8th edition of the TNM system.	[VALUE CHANGE] 8th edition to be added (2017).	Add 8th revision to MDR. [DONE] Answer to Frankfurt (Gabriele Husman) [DONE]	We will add the 8th edition to the TNM system.

7	BRAF_PIC3CA_HER_MUTATION_STATUS	LIST_OF_VALUES [BRAF, PIC3CA. HER2 mutation status - not mutated, BRAF, PIC3CA. HER2 mutation status - mutated, BRAF, PIC3CA. HER2 mutation status - not done]	If only one (or 2) of the 3 mutation analyses has / have been carried out, does this mean "not done"?	[COMMENT CHANGE] If only 1 or 2 of the three mutation analyses have been done, the "Partial information available" value shall be selected. [VALUE ADDITION] "Partial information available"	Update MDR. [DONE] Answer to Frankfurt (Gabriele Husman) [DONE]	We will add a value "partial information available" to the data model, so In this case you can select that.
8	AGE_AT_PRIMARY_DIAGNOSIS		Is it possible to calculate the age at primary diagnosis from the entries in the database (from date of primary diagnosis and date of birth) or must it always be entered manually?	[NOTHING]	[Nothing.] Answer to Frankfurt (Gabriele Husman) [DONE]	There is **no date of birth collected** in the data model. The only age at primary diagnosis is directly collected, rounded to years. (approach taken to minimize the likelihood of patient re-identification). It is the responsibility of the biobank to calculate the at the primary diagnosis based on data in their source systems (possibly containing absolute timestamps of events and absolute dates) and provide it into the CCDC already as age at primary diagnosis (which is relative). This is part of the data transformation process.
9	TIME_OF_RECURRENCE_RELATIVE	Time of recurrence (**metastasis** diagnosis)	1) Is it possible to calculate the period of time between primary diagnosis and time of recurrence from the entries in the database or must it always be entered manually? 2) What is meant by "Time of recurrence"? Distant metastasis only or local recurrence and / or newly	[COMMENT CHANGE] Any re-occurrence of cancer, be it a local re-occurrence, a lymph node metastasis, or a distant metastasis.	MDR update [DONE] Answer to Frankfurt (Gabriele Husman) [DONE]	1) There is **no date of primary diagnosis** collected, only the age at primary diagnosis (approach taken to minimize the likelihood of patient re-identification). It is the responsibility of the biobank to calculate the at the primary diagnosis based on data in their source systems (possibly containing absolute timestamps of events and absolute dates)

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

			diagnosed lymph node metastasis as well?			and provide it into the CCDC already as age at primary diagnosis (which is relative). This is part of the data transformation process. 2) It means re-occurrence of cancer, be it a local re-occurrence, a lymph node metastasis, or a distant metastasis.
10	PHARMACOTHERAPY_SCHEME		What should be documented if a patient did not receive a complete chemotherapy (for example stop of chemotherapy or dosage reduction due to adverse effects)?	[COMMENT CHANGE] If the therapy was terminated or changed (e.g., dosage reduced), “Other” shall be selected. Additional textual information should be provided in such a case, see [Item 14].	Add the clarification into MDR to use “Other” for cases where the therapy did not follow the procedure. Answer to Frankfurt (Gabriele Husman) [DONE]	If the theraphy was terminated or changed (e.g., dosage reduced), “Other” shall be selected. There will be a plain text field to be used in these cases to provide this information.
11	RADIATION_THERAPY_START_RELATIVE		How should we document a combined radio-chemotherapy? 2 entries?	[COMMENT CHANGE] For combined therapies, they should be entered as separate therapies. They can be correlated later because start time is provided (relative to the primary diagnosis - i.e., this attribute *_START_RELATIVE) and it is obvious that combined therapies start at the same time.	Add clarification into MDR. Answer to Frankfurt (Gabriele Husman) [DONE]	For combined therapies, they should be entered as separate therapies.

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

12	HIST_MORPHOLOGY	LIST_OF_VALUES [Other, Undifferentiated carcinoma, Mixed adenoneuroendocrine carcinoma, Spindle cell carcinoma, Serrated adenocarcinoma, Micropapillary carcinoma, Adeonsquamous carcinoma, Squamous cell carcinoma, small cell neuroendocrine carcinoma, Large cell neuroendocrine carcinoma, High-grade neuroendocrine carcinoma, Medullary carcinoma, Signet-ring cell carcinoma, Mucinous carcinoma, Adenocarcinoma]	How important are the subtypes? Now it seems there are too many different options. Most complicated value to the biobanks.	[COMMENT CHANGE] 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. [CONSISTENCY CHECK] Adenocarcinoma precludes combination with G4 grade => it does not describe all the cases that might be collected.	MDR update. [DONE] Correct the answer to Uppsala (Per-Henrik)	This item 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.
13	WHO_GRADE	LIST_OF_VALUES [WHO Grading - Grade - G4, WHO Grading - Grade - G3, WHO Grading - Grade - G2, WHO Grading - Grade - G1]	How to map the grading to WHO Grades e.g. in case of Sweden where there are only two grades used (medium high and low medium). Swedish approach: The two grades is the standard for CRC, including surgical resections of primary tumors in Sweden. (2-level answers). Statistically, most low/mod cases would correspond to G2	[COMMENT CHANGE] For Sweden "medium high" shall map to G3, and "low medium" shall map to G2. This **has to be documented** in the provenance information.	MDR update [DONE] This has been communicated to Uppsala (Per-Henrik).[DONE]	For Sweden "medium high" shall map to G3, and "low medium" shall map to G2. This **has to be documented** in the provenance information.

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

			and most high/mod cases would correspond to G.			
14	PHARMACOTHERAPY_SCHEME	LIST_OF_VALUES [Other, Scheme of pharmacotherapy - 5-FU 325-350 mg/m2 + LV 20 mg/m2i.v. bolus, day1-5, weeks 1 and 5, Scheme of pharmacotherapy - 5-FU 400 mg/m2 + 100 mg i.v. bolus, d 1,2, 11,12,21,22, Scheme of pharmacotherapy - 5-FU 225 mg/m2 i.v. continuous infusion, 5 days per week, Scheme of pharmacotherapy - 5 -FU 1000 mg/m2 i.v. continuous infusion, day 1-5, weeks 1 and 5, Scheme of pharmacotherapy - Capecitabine 800-825 mg/m2 bid po, day 1-5, together with radiation or continuously untill end of radiation, Scheme of pharmacotherapy - UFT (300-340mg/m2/day) and LV (22.5-90 mg/day) po continuously, 5(-7) days per week, together	Are the options too specific?	[ATTRIBUTE ADDITION] When Other option is selected for PHARMACOTHERAPY_SCHEME, the plain text description shall be provided (new PHARMACOTHERAPY_SCHEME_DESCRIPTION attribute). The plain text must include at least the chemical compounds used, the dosage and timing is optional. [COMMENT CHANGE] When Other option is selected for PHARMACOTHERAPY_SCHEME, the plain text description into PHARMACOTHERAPY_SCHEME_DESCRIPTION should be provided. [IMPLEMENTATION NOTE] This attribute has to support copy-paste to copy it across multiple patients in the manual data entry interface.	Update MDR. Notify Uppsala (Per-Henrik). This was the decision of the panel of medical research experts designing the data model, based on common clinical protocols. There is always an option of “other” if the treatment has not followed one of those protocols.	When Other option is selected for PHARMACOTHERAPY_SCHEME, the plain text description shall be provided. A new PHARMACOTHERAPY_SCHEME_DESCRIPTION attribute will be added. The plain text must include at least the chemical compounds used, the dosage and timing is optional.

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

		with radiotherapy, Scheme of pharmacotherapy - 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]				
15	DIAG_CT_DONE DIAG_MRI_DONE DIAG_X_DONE DIAG_ Liver imaging _DONE DIAG_COLONOSCOPY		Are these values meant as whether the method has been used as a part of the initial diagnosis or as any follow up? Does the "data available" mean that it is directly provided as a part of the data set?	[COMMENT CHANGE] These values shall be TRUE only if they were done within the context of the primary diagnosis.	They are meant if they were used for the initial diagnosis. Add information into the MDR that 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.	These values shall be selected only if they were done within the context of the primary diagnosis.
16	Definition of Surgical tumor (is biopsy ok?)			[COMMENT ADDITION TO INCLUSION CRITERIA] Biopsies do not qualify as surgical tumor material. Biopsies do not provide sufficient amount of material to support multiple research projects.	Answer to Helsinki Biobank (Olli Carpen).	Biopsies do not qualify as surgical tumor material. Biopsies do not provide sufficient amount of material to support multiple research projects.

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

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

17	All variables in Diagnostic Exam category (DIAG_CT_DONE etc.)		Do we have to deliver any further information like images or medical results of the investigation?	[COMMENT CHANGE] Data itself is not going to be collected as a part of the central data set. This acts as advertisement to the researchers to ask for more data if they find relevant cases This depends on your answer. If you have 'data available' it may be requested at some point by researchers. Any further information may also be requested by researchers but not collected by BBMRI-ERIC.	Answer to Wurzburg (file ADOPT\WP3\Qs_Wurzburg) [DONE]	Data itself is not going to be collected as a part of the central data set. This acts as advertisement to the researchers to ask for more data if they find relevant cases
18	DIGITAL_IMAGING_AVAILABILITY & DIGITAL_IMAGING_INVASION_FRONT		Will images be accepted that don't have a magnification of 40x ?	[COMMENT]: This issues is of later decision and later phase of the project (if the 40x is needed or both options are allowed). [COMMENT] Resolutions should be <0.125um/pixel (this is more accurate description of 40x). [NOTE] Assuming <6GB/file, 10,000 cases results <60TB. Assuming <1,000 cases per biobank, 6TB can be shipped on a hard drive in the worst case.	Answer to Wurzburg (file ADOPT\WP3\Qs_Wurzburg) [DONE]	This issues is of later decision and later phase of the project. At stage we collect the availability of 40x (<0.125um/pixel).
19	HIST_METASTASIS		Metastasis can differ between stages of the cancer disease. From which stage of the disease do we need the information about metastasis?	[COMMENT] This is intended for primary diagnosis only.	Answer to Wurzburg (file ADOPT\WP3\Qs_Wurzburg) [DONE]	This is intended for primary diagnosis only.

				[ATTRIBUTE ADDITION] BIOLOGICAL _MATERIAL_FROM_RECURRENCE_ AVAILABLE (BOOLEAN, OPTIONAL) - only if TIME_OF_RECURRENCE_RELATIVE is set.		
20	UICC_STAGE		Is it possible to have an unknown UICC Stage?	[COMMENT] Stage is essential, no.	Answer to Wurzburg (file ADOPT\WP3\Qs_Wurzburg) [DONE]	Stage is essential, so it is not possible to have unknown as a value.
21	MM_MISMATCH_REPAIR_GE		After a consultation with our oncology department it's unclear which marker is meant. Are there any further explanations?	[NOTHING] [COMMENT] This should be clear based on the data model. In column E explains what is meant exactly.	Answer to Wurzburg (file ADOPT\WP3\Qs_Wurzburg) [DONE]	Column E explains what is meant exactly.
22	MM_RISK_SITUATION_HNPCC		We need further explanation of this item	[NOTHING] [COMMENT]: This should also be clear based on the data model: in line 32 how the statement is generated + including the link to literature.	Answer to Wurzburg (file ADOPT\WP3\Qs_Wurzburg). [DONE]	It is explained in line 32 how the statement is generated including the link to literature.
23	DATE_DIAGNOSIS		For us the start date is the day of surgery. This is also the date we use to calculate the AGE_AT_PRIMARY_DIAGNOSIS. For those patients who had a colonoscopy done at our hospital we could use the day of	[COMMENT CHANGE] Histopathological diagnosis by biopsy or surgery qualifies as primary diagnosis.	Update MDR. [DONE] Answer to Lubeck (Lars Boekmann/ Jens Haberman)	Use the surgery date as primary diagnosis date.

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

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

			colonoscopy as the start date. But for those who had the colonoscopy done at external practices we don't have the information of the exact date.		[DONE]	
24	DIAG_COLONOSCOPY		There are only three values for this parameter (not done, positive, negative). If colonoscopy was performed externally we can't tell if it was positive or negative. Hence, in such cases we select "not done".	[COMMENT CHANGE] 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".	Update MDR. [DONE] Answer to Lubeck (Lars Boekmann/ Jens Haberman) [DONE]	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"
25	MM_MICROSAT_INSTABILITY		At the time when most of the patients have been recruited, MSI-status and other molecular markers were not yet analyzed routinely. But we have performed DNA image cytometry for all the patients to determine the ploidy status. This is valuable information too. Maybe this could be included?	[COMMENT CHANGE] Image cytometry does **not** qualify for comparability reasons.	Update MDR. [DONE] Answer to Lubeck (Lars Boekmann/ Jens Haberman) [DONE]	Image cytometry does **not** qualify for MSI satus or other molecular markers status for comparability reasons. We are not extending the data model at this point in order to minimize the disruptions for the biobanks that are already preparing the datasets.
26	PHARMACOTHERAY_SCHEME		The list of values is very specific for very few therapy schemes. But there are many other therapy regimens including e. g. oxalipaltin or irinotecan. If we would only use those values listed in list of values than this field will be empty for almost all patients.	[DUPLICATE OF 14]	Answer to Lubeck (Lars Boekmann/ Jens Haberman) [DONE]	In this case select 'other' option. When Other option is selected for PHARMACOTHERAPY_SCHEME, the plain text description shall be provided. A new PHARMACOTHERAPY_SCHEME_DESCRIPTION attribute will be added. The plain text must include at least the chemical compounds used, the dosage and timing is optional.

27	DIAGNOSTIC IMAGING		Diagnostic imaging is everything that is performed was performed before surgery?	[COMMENT CHANGE] It is imaging that has been done in the context the primary diagnosis. It can be before the surgery or shortly after the surgery, depending on the medical workflow.	Answer to Lubeck (Lars Boekmann/ Jens Haberman) [DONE]	It is imaging that has been done in the context the primary diagnosis. It can be before the surgery or shortly after the surgery, depending on the medical workflow.
28	Surgery		Is it true that we need only surgically treated cancer cases, i.e. no pre-operatively treated cases?	[PARTIAL DUPLICATE OF 16] [COMMENT CHANGE] 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.	Answer to Gewebe biobank (Inti Zlobec) [DONE]	Pre-operatively treated cases (neoadjuvant therapy) are welcome, but there needs to be surgery later on anyways, to have also sufficient amount of biological material.
29	SURGERY_TYPE		Why is tumor location (right/left/rectum, for e.g.) not included? Tumor location could be relevant for several situations and is a known prognostic factor	[VALUE CHECK] Check that we have all the possible locations in the list of permitted values: C18.0 Cecum, C18.1 Appendix, C18.2 Ascending (right) colon, C18.3 Hepatic flexure, C18.4 Transverse colon, C18.5 Splenic flexure, C18.6 Descending (left) colon, C18.7 Sigmoid, C19 Rectosigmoid, C19.9 Rectosigmoid, C20 Rectum, C20.9 Rectum	Update MDR. [DONE] Answer to GEwebe biobank (Inti Zlobec) [DONE]	The HIST_LOCALIZATION is already included as a required attribute.
30	Surgery		Should all the surgeries be included? Surgery can be done depending on the prognosis of a patient.	[COMMENT] Sometimes excision is done, not curative surgery - e.g., when there is distant metastases in other organs. For the purpose of the CRC-Cohort, other types of	Answer to Malta (Malcolm Pace) [DONE]	

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

				surgery are also accepted (e.g., palliative surgery).		
31	Surgery		Is it only the first surgery that the patient had prior to diagnosis, has to be included? What is the cut off point because a patient can still have therapy scheduled for this year or the next and we cannot stay updating every case, it will take us forever.	[COMMENT] Multiple surgeries are supported, if available. Multiple treatments (be it surgeries or other types of treatments) can be added - there is 1..N relation between treatment and patient. Multiple surgeries per patient draw quite some interest by researchers.	Answer to Malta (Malcolm Pace)[DONE]	

30 a) Is the sample age important? - No (is this correct?) [ATTRIBUTE ADDITION] SAMPLE_COLLECTION_YEAR (Year when the sample was retrieved.)

30 b) Can a biopsy be used as a sample? No (is this correct?) Correct [DUPLICATE OF 16]

#	Data Item	List of values	Question	Comments	Action
32	DIAG_X_DONE (Lung imaging)	LIST_OF_VALUES [Lung imaging - Unknown, Lung imaging - Done, data not available, Lung imaging - Done, data available, Lung imaging - Not done]	Does this include lung X-ray only, or CT and MRI of the lung as well?	[COMMENT CHANGE] If CT or MRI or PET scan is available, this should be also considered one of the "Done" options.	Update MDR. [DONE] Answer to Frankfurt (Gabrielle Husman) [DONE]

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

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

33	DIGITAL_IMAGING_AVAILABILITY	LIST_OF_VALUES [No, Can be generated, Readily available]	Does this apply only to (larger) specimen obtained during surgery, or for biopsies as well?	[COMMENT CHANGE] 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.	Update MDR. [DONE] Need to specify this Answer to Frankfurt (Gabrielle Husman) 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.
34	DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY	LIST_OF_VALUES [Invasion front not included, Readily available, Can be generated, No]	If there are only specimen from a biopsy available, should we document “Invasion front not included” in this case?	[DUPLICATE 33] Imaging of biopsy itself does not qualify.	Answer to Frankfurt (Gabrielle Husman) [DONE]
35	TNM_DISTANT_METASTASIS	LIST_OF_VALUES [Distant metastasis - M1b, Distant metastasis - M1a, Distant metastasis - M1, Distant metastasis - M0]	**M1c** is included in the new version of TNM (8th edition). Please add in the BBMRI database. **MX** to be added (means M was not evaluation.	[VALUE ADDITION] M1c and MX to be added.	Update MDR. [DONE] The new version of TNM 8th edition need to be included in the BBMRI database Answer to Frankfurt (Gabrielle Husman) [DONE]
36	UICC_STAGE	LIST_OF_VALUES [Stage - IVB, Stage - IVA, Stage - IIIC, Stage - IIIB, Stage - IIIA , Stage - IIB,	**IVC** is included in the new version of TNM (8th edition). Please add in the BBMRI database.	[VALUE ADDITION] 8th edition	Update MDR. [DONE]

		Stage - II A, Stage - I, Stage - 0]			
37	WHO_GRADE	LIST_OF_VALUES [WHO Grading - Grade - G4, WHO Grading - Grade - G3, WHO Grading - Grade - G2, WHO Grading - Grade - G1]	**GX** (Grade cannot be assessed) is missing	[VALUE ADDITION] GX	Update MDR. [DONE]
38	TNM_PRIMARY_TUMOR	LIST_OF_VALUES [Primary Tumor - T4b, Primary Tumor - T4a, Primary Tumor - T3, Primary Tumor - T2, Primary Tumor - T1, Primary Tumor - Tis, Primary Tumor - T0, Primary Tumor - TX]	What if there is only T4 and not the differentiation between T4a and T4b documented in the biobank data set?	[VALUE ADDITION] T4 should be added as is.	Update MDR. [DONE]
39	TNM_REGIONAL_LYMPH_NODES	LIST_OF_VALUES [Regional lymph nodes - N2b, Regional lymph nodes - N2a, Regional lymph nodes - N2, Regional lymph	N3 is missing (might be a question of occurence)	[VALUE ADDITION] N3 should be added.	Update MDR. [DONE]

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

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

		nodes - N1c, Regional lymph nodes - N1b, Regional lymph nodes - N1a, Regional lymph nodes - N1, Regional lymph nodes - N0, Regional lymph nodes - NX]			
40	SAMPLE_MATERIAL_TYPE	LIST_OF_VALUES [Other, Healthy colon tissue, Tumor]	should material types like DNA, blood etc. be mapped to Other? Or is this not relevant for colon cancer	It should be mapped to others.	
41	PHARMACOTHERAPY_SCHEME		1) Is the list of values a standardized list? What is the standard? 2) How to deal with unknown schemes?		Yes, please see article https://academic.oup.com/annonc/article/23/10/2479/195121 Table 11. In unknown schemes you can select "other" and provide in the text field the explanation (newly introduced attribute PHARMACOTHERAPY_SCHEME_DESC), see answer to #14 . Answer to Wursburg [DONE]

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

42	RADIATION_THERAPY_END_RELATIVE		How to deal with several therapies?		They should be entered as separate therapies identified by their relative start date (relative to primary diagnosis, as all other relative dates in the data model). This also holds true for combined therapies.. Answer to Wursburg [DONE]
43	THERAPY_RESPONSE_TIMESTAMP_RELATIVE		All Follow-Ups ?		If you have them, yes - you can have multiple therapy responses per patient, so multiple can be provided. Answer to Wursburg [DONE]
44	SAMPLE_ID		All samples ? Don't you need sample collection date/time ?		You can include all samples you have (inclusion criteria is FFPE) - for each you should provide a preservation mode and material type (see SAMPLE_MATERIAL_TYPE and SAMPLE_PRESERVATION_MODE attributes). We will be collecting a sample acquisition year (this will be a newly introduced attribute). Answer to Wursburg [DONE]
45	SURGERY_RADICALITY (+ other variables in Surgery)		How to deal with several surgeries? // How to deal with unknown values?		Multiple surgeries can be provided. Sometimes excision is done, not curative surgery, e.g. when there is distant metastases in other organs. For the CRC cohort, other types of surgery are also accepted (e.g. pallative surgery) Answer to Wursburg [DONE]

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

46	TARGETED_THERAPY_END_RELATIVE		After consultation with our oncological department it's still unclear what a targated therapie is. Are there any further explanations? How to deal with unknown dates?		Pallative surgery is an example of targeted therapy. If start date is unknown, you cannot include them (required attribute). If end date is unknown, it is no problem (optional attribute).
47	OVERALL_SURVIVAL_STATUS		Shouldn't this be weeks after diagnosis ?		Because the overall survival is normally counted after removal of the tumor, i.e., after first surgical treatment. Answer to Wursburg [DONE]
48	PHARMACOTHERAPY_END_RELATIVE		In older cases exact dates were not always documented. However, it is documented that a pharmacotherapy took place. Suggestion: We could either only save pharmacotherapy schemes (or substances) or a default value in pharmacotherapy start and end (something like "99") to indicate that the patient received a pharmacotherapy. Additionally, after consultation with our onkology department we are not sure, if dates itself are relevant here. The statement if a pharmacotherapy took place before the sample was taken might be more relevant to researchers.		This is a required field so in case of missing dates, the older cases should not be included. However, we are not collecting absolute dates but relative dates rounded to weeks. So if you are able to infer the timing with this precision, it would be fine. Answer to Wursburg (new Qs: Qs Wursburg 2.xls) [DONE]
49	PHARMACOTHERAPY_START_RELATIVE		In older cases exact dates were not always documented. However, it is documented that a pharmacotherapy took place. Suggestion: We could either only save pharmacotherapy schemes (or substances) or a default value in pharmacotherapy start and end (something like "99") to indicate that the patient received a pharmacotherapy		This is a required field so in case of missing dates, the older cases should not be included. However, we are not collecting absolute dates but relative dates rounded to weeks. So if you are able to infer the timing with this precision, it would be fine

					Answer to Wursburg (new Qs: Qs Wursburg 2.xls) [DONE]
50	PHARMACOTHERAPY_SCHEME		After another consultation with our onkology department we are not sure if this list of values makes sense. In general, there are 3 main substances that might be given in hundreds of different combinations and doses. Additionally, some biological substances might be involved as well. When we have to map our schemes to the present ones, we would need to chose "Other" most of the time, although we might have used some described combinations but with slightly different doses etc. . Propably it would make sense to have a field for "substances" and one for "combination" and one for "doses" or to have a LOV that includes a whole common schemata catalogue.		Please note that the list of values comes from this: https://academic.oup.com/annonc/article/23/10/2479/195121 Table 11. When Other option is selected for PHARMACOTHERAPY_SCHEME, the plain text description shall be provided. A new PHARMACOTHERAPY_SCHEME_DESCRIPTION attribute will be added. The plain text must include at least the chemical compounds used, the dosage and timing is optional. Duplicate to #14 & #41 Answer to Wursburg (new Qs: Qs Wursburg 2.xls) [DONE]
51	THERAPY_RESPONSE		1) We also do have the case "Unknown response", should we add it to the List of values? 2) By "therapy" do you mean therapy including or exluding surgeries? (In colon cancer, surgeries are a common part of therapy block 1.) 3) The LOV has some mixed wordings: "response" describes the transformation of a tumor under therapy whereas "stable desease" for example is more common in contexts of status of desease (not particularly in contexts of a tumor itself). Should "complete response" rather mean "complete		1) We will add “unknown” to the list of values. 2) Yes, surgery is also a form of therapy 3) Please note that the list of values comes from RECIST criteria: https://ctep.cancer.gov/protocoldevelopment/docs/recist_guideline.pdf Answer to Wursburg (new Qs: Qs Wursburg 2.xls) [DONE]

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

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

			remission" and "partial response" "partial remission" ?		
52	TARGETED_THERAPY_END_RELATIVE		We believe that targeted therapy are therapies according to https://www.cancer.gov/about-cancer/treatment/types/tar geted-therapies/targeted-therapies-fact-sheet?		This is ok, we do not specify here the targeted therapies. You may use the targeted therapies according to the NCI. Answer to Wursburg (new Qs: Qs Wursburg 2.xls) [DONE]
53	TARGETED_THERAPY_START_RELATIVE		We believe that targeted therapy are therapies according to https://www.cancer.gov/about-cancer/treatment/types/tar geted-therapies/targeted-therapies-fact-sheet How to deal with unknown dates?		The same as above. This is a required field so in case of missing dates, the cases should not be included. However, again, we are not collecting absolute dates but relative dates rounded to weeks. So if you are able to infer the timing with this precision, it would be fine. Answer to Wursburg (new Qs: Qs Wursburg 2.xls) [DONE]
54	DIAG_ Liver imaging _DONE		there is option: Liver imaging – Unknown and Liver imaging - Unknown Not done, it is not consistent with previous fields		Yes, it is a typo, we will fix it. Thank you for noticing. Answer to Bern (Magdalena) [DONE]
55	DIAG_COLONOSCOPY		also not consistent with previous once, Lacking Colonoscopy diagnostic exam- Unknown or Done, data not available		As colonoscopy does not result in additional data, it is not collected in terms of data available/ not available. We will add unknown value to the list of values. Answer to Bern (Magdalena) [DONE]

56	TARGETED_THERAPY_START_RELATIVE		In is required field, is there going to be option for unknown, or not applicable?		Relative date is only required if the therapy occurred so when there is not therapy it is not applicable and thus ok.
57	PHARMACOTHERAPY_END_RELATIVE , PHARMACOTHERAPY_START_RELATIVE , SURGERY_START_RELATIVE , RADIATION_THERAPY_END_RELATIVE, RADIATION_THERAPY_START_RELATIVE , THERAPY_RESPONSE_TIMESTAMP_RELATIVE , THERAPY_RESPONSE		would there be option Unknown/data not available? In our case we do not have information/full information about those fields, if there would not be option, should be leave those fields blank?		These data fields are only required if pharmacotherapy took place. This holds true for any therapies. When Other option is selected for PHARMACOTHERAPY_SCHEME, the plain text description shall be provided. A new PHARMACOTHERAPY_SCHEME_DESCRIPTION attribute will be added. The plain text must include at least the chemical compounds used, the dosage and timing is optional. Duplicate to #14 & #41 Answer to Bern (Magdalena) [DONE]

One biobank is requesting the list of morphology codes eligible for this study - what does this mean?

Survival:

- Disease free survival is counted really from the initial treatment and makes only sense for **treated *and* localized** tumors, that can be completely removed (for metastases left inside the body, you can't call it "disease free").

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

[CONCLUSION] *We are not collecting disease free survival explicitly. Extending the data model substantially at this point is really not desirable. Some estimates can be obtained from TIME_OF_RECURRENCE_RELATIVE.*

- There is, however, one problem - and I need to call Kurt again: we are collecting “Overall survival” and not “Disease free survival” - I was confused from the call with Per-Hendrik, when he was speaking about disease-free survival. There are three possible solutions:
    - Overall survival = Disease-free survival [in medicine]
    - We leave overall survival and switch the time anchor to the diagnosis.
    - We switch to disease-free survival and let it anchored to the treatment.
- **Resolution from second call with Kurt:**
    - The Overall survival should be kept as Overall - and anchored relative to the initial diagnosis (as all other relative dates are). It is also more reliable value and should be available for most patients.
    - The Disease-free survival should be added as OPTIONAL value - and should be anchored relative to the treatment (and should be explicitly commented why).

Schema for CCDC (Qs from Sebastian):

- Location of metastasis permitting values of include the string “localization of metastasis” - this is not a clean definition
- Inclusion of “unknown’ as an data value for se

CRC cohort - technical comments to MDR / XDR

Coding of values

- coding of (enum) values should be consistent, especially for NULL values
- is it necessary to prefix the the codes with the record name (it’s also not done consistent, e.g. surgery and vital status don't use a prefix)

Null values

- a null value of “Not Specified” would allow to store intermediate Records, if the value s REQUIRED
- there should be a distinction between “not applicable”, “not known”, “not specified”
- the semantics of the “Locations, BasicData, LongititinalData, Events ...” type should be explained, it only appears in the XSD
- in the example “name” is always a date, should it be interpreted as date
- is there a definition how often an event can occur 0/1/*

Events

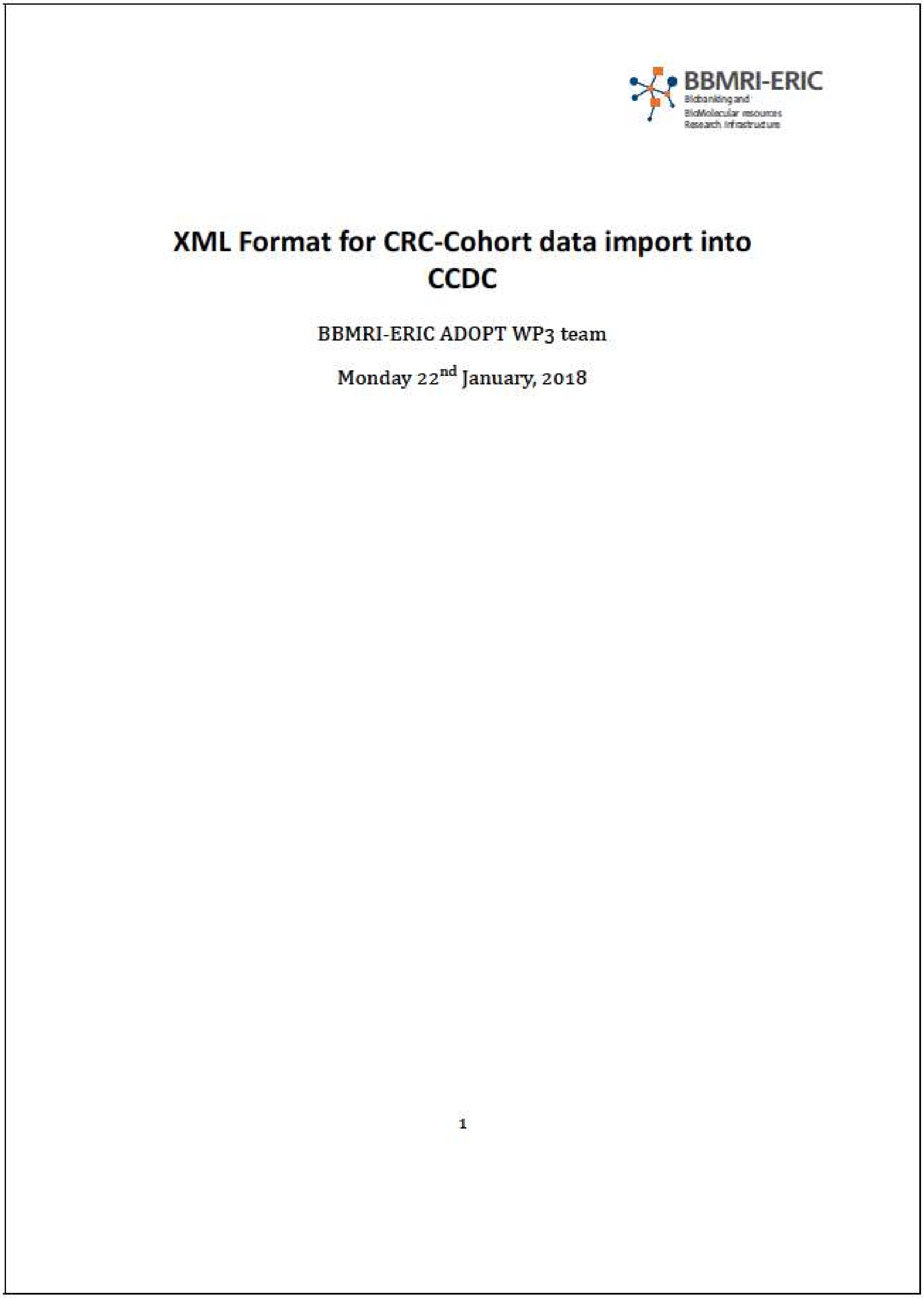
- ● the semantics of the “Locations, BasicData, LongititinalData, Events ...” type should be explained, it only appears in the XSD
- ● in the example “name” is always a date, should it be interpreted as date
- ● is there a definition how often an event can occur 0/1/*

Sample

- is “sample” per definition a stored sample in the biobank
- do we only cover sample of colon, i.e. not liver/lymph nodes/etc.

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

B. Instructions for Biobanks



BBMRI-ERIC Biobanking and BioMolecular resources Research Infrastructure

XML Format for CRC-Cohort data import into CCDC

BBMRI-ERIC ADOPT WP3 team

Monday 22nd January, 2018

1

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

Contents

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

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

1 Introduction

This document provides documentation on how to construct the XML file to import the data for the CRC-Cohort contributors into the CCDC service. It is generated based on the XSD definition, which defines constrains on the XML to be imported into the CCDC.

Please note that using XML it is possible to import incomplete data, to allow for semi-automated import process:

1. the data, which biobanks have already structured, are imported via XML import,

z. each individual patient case is manually completed using CCDC.

For this reason, the XML import does not require all the data elements marked as REQUIRED in the data model. In the manual editing using CCDC web interface, before each patient can be saved, completeness of all the REQUIRED fields is checked.

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

2 Patient Pseudonyms

Pseudonyms of patients must be generated in compliance with the Data Protection Policy of CRC-Cohort. They are written into `<Identifier>...</Identifier>` element of the XML, as demonstrated in the example XML.

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

3 Forms

3.1 form_28_ver-27 - Form

Required data elements

- Dataelement_14_3 - MM_MICROSAT_INSTABILITY
- Dataelement_15_z - MM_MISMATCH_REPAIR_GE
- Dataelement_z0_3 - MM_KRAS_MUTATION_KRAS_EXz
- Dataelement_z1_5 - MM_KRAS_MUTATION_KRAS_EX3
- Dataelement_zz_4 - MM_KRAS_MUTATION_KRAS_EX4
- Dataelement_z3_5 - MM_KRAS_MUTATION_NRAS_EXz
- Dataelement_z4_4 - MM_KRAS_MUTATION_NRAS_EX3
- Dataelement_z5_3 - MM_KRAS_MUTATION_NRAS_EX4
- Dataelement_30_3 - DIAG_MRI_DONE
- Dataelement_31_3 - DIAG_CT_DONE
- Dataelement_3_1 - AGE_AT_PRIMARY_DIAGNOSIS
- Dataelement_5_z - VITAL_STATUS
- Dataelement_61_5 - DIAG_LIVER_IMAGING_DONE
- Dataelement_63_4 - DIAG_X_DONE
- Dataelement_7_z - OVERALL_SURVIVAL_STATUS
- Dataelement_85_1 - SEX
- Dataelement_88_1 - DIAG_COLONOSCOPY

Optional data elements

- Dataelement_16_3 - MM_RISK_SITUATION_HNPCC
- Dataelement_z_z - CLINICAL_STUDY_PARTICIPANT
- Dataelement_4_3 - TIME_OF_RECURRENCE_RELATIVE
- Dataelement_51_3 - DATE_DIAGNOSIS
- Dataelement_87_1 - BRAF_PIC3CA_HER_MUTATION_STATUS

Other data elements

- Dataelement_6_3 - VITAL_STATUS_TIMESTAMP
  Level: `if(VITAL_STATUS!=UNKNOWN){REQUIRED}else{OPTIONAL}`

3.2 form_29_ver-5 - Form5

Required data elements

- Dataelement_1z_4 - RADIATION_THERAPY_START_RELATIVE
- Dataelement_13_z - RADIATION_THERAPY_END_RELATIVE

Optional data elements None.

Other data elements None.

3.3 form_30_ver-3 - Form6

Required data elements

- Dataelement_35_3 - TARGETED_THERAPY_START_RELATIVE

Optional data elements

- Dataelement_36_1 - TARGETED_THERAPY_END_RELATIVE

Other data elements None.

3.4 form_31_ver-2 - Form4

Required data elements

- Dataelement_33_1 - THERAPY_RESPONSE
- Dataelement_34_1 - THERAPY_RESPONSE_TIMESTAMP_RELATIVE

Optional data elements None.

Other data elements None.

3.5 form_32_ver-8 - Form

Required data elements

- Dataelement_49_1 - SURGERY_TYPE
- Dataelement_8_3 - SURGERY_START_RELATIVE
- Dataelement_93_1 - SURGERY_LOCATION
- Dataelement_9_z - SURGERY_RADICALITY

Optional data elements

- Dataelement_67_1 - SURGERY_TYPE_OTHER

Other data elements None.

3.6 form_33_ver-10 - Form3

Required data elements

- Dataelement_10_z - PHARMACOTHERAPY_START_RELATIVE
- Dataelement_11_z - PHARMACOTHERAPY_END_RELATIVE
- Dataelement_59_5 - PHARMACOTHERAPY_SCHEME

Optional data elements None.

Other data elements

- Dataelement_81_3 - PHARMACOTHERAPY_SCHEME_DESCRIPTION
  Level: `if(PHARMACOTHERAPY_SCHEME==Other){REQUIRED}else{OPTIONAL}`

3.7 form_34_ver-22 - Form2

Required data elements

- Dataelement_53_3 - WHO_GRADE_VERSION
- Dataelement_68_z-HIST_METASTASIS
- Dataelement_70_z-UICC_STAGE
- Dataelement_71_1 - TNM_PRIMARY_TUMOR
- Dataelement_73_3 - UICC_VERSION

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

- Dataelement_75_1 - TNM_DISTANT_METASTASIS
- Dataelement_77_1 - TNM_REGIONAL_LYMPH_NODES
- Dataelement_83_1 - WHO_GRADE
- Dataelement_91_1 - HIST_MORPHOLOGY
- Dataelement_9z_1 - HIST_LOCALIZATION

Optional data elements

- Dataelement_57_3 - DIGITAL_IMAGING_AVAILABILITY
- Dataelement_58_z - DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY
- Dataelement_8z_1 - BIOLOGICAL_MATERIAL_FROM_RECURRENCE_AVAILABLE

Other data elements None.

3.8 form_35_ver-6 - Form1

Required data elements

- Dataelement_54_z - SAMPLE_MATERIAL_TYPE
- Dataelement_55_z - SAMPLE_PRESERVATION_MODE
- Dataelement_56_z-SAMPLE_ID
- Dataelement_89_3 - YEAR_OF_SAMPLE_COLLECTION

Optional data elements None.

Other data elements None.

4 Data Elements

This section provides documentation of individual data elements, based on CRC-Cohort data model and XSD.

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

4.1 Dataelement_10_2 - PHARMACOTHERAPY_START_RELATIVE

XSD label Dataelement_10_z

Data model label PHARMACOTHERAPY_START_RELATIVE

Level in data model `REQUIRED`

XSD name Date of start of pharamacotherapy

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [week]   (0<=x)
```

XSD parent form Form3 - form_33_ver-10

XSD description
Start of the drug intake in weeks since initial diagnosis.

Data model description
Start of the drug intake in weeks since initial diagnosis.

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

4.2 Dataelement_11_2 - PHARMACOTHERAPY_END_RELATIVE

XSD label Dataelement_11_z

Data model label PHARMACOTHERAPY_END_RELATIVE

Level in data model REQUIRED

XSD name Date of end of pharamcotherapy

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [week]  (0<=x)
```

XSD parent form Form3 - form_33_ver-10

XSD description

End of the drug intake in weeks since initial diagnosis.

Data model description

End of the drug intake in weeks since initial diagnosis.

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

4.3 Dataelement_12_4 - RADIATION_THERAPY_START_RELATIVE

XSD label Dataelement_1z_4

Data model label RADIATION_THERAPY_START_RELATIVE

Level in data model `REQUIRED`

XSD name Date of start of radiation therapy

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [week]   (0<=x)
```

XSD parent form Form5 - form_z9_ver-5

XSD description

Start of the radiation therapy in weeks since initial diagnosis. For combined therapies, they should be entered as separate therapies.

Data model description

Start of the radiation therapy in weeks since initial diagnosis. For combined therapies, they should be entered as separate therapies.

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

4.4 Dataelement_13_2 - RADIATION_THERAPY_END_RELATIVE

XSD label Dataelement_13_z

Data model label RADIATION_THERAPY_END_RELATIVE

Level in data model `REQUIRED`

XSD name Date of end of radiation therapy

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [week]  (0<=x)
```

XSD parent form Form5 - form_z9_ver-5

XSD description
End of the radiation therapy in weeks since initial diagnosis.

Data model description
End of the radiation therapy in weeks since initial diagnosis.

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

4.5 Dataelement_14_3 - MM_MICROSAT_INSTABILITY

XSD label Dataelement_14_3

Data model label MM_MICROSAT_INSTABILITY

Level in data model `REQUIRED`

XSD name Microsatellite instability

XSD type xs:string

List of permitted values in XSD

```
”YES”
”NO”
”NOT_DONE”
```

Type in data model

```
LIST_OF_VALUES [no; yes; not done]
```

XSD parent form Form - form_z8_ver-z7

XSD description

Microsatellites analysed BATz6, D17Sz50, D5S346, BAT40, DzS1z3 and BATz5. Image cytometry does not qualify for comparability reasons

Data model description

Microsatellites analysed BATz6, D17Sz50, D5S346, BAT40, DzS1z3 and BATz5. Image cytometry does not qualify for comparability reasons

4.6 Dataelement_15_2 - MM_MISMATCH_REPAIR_GE

XSD label Dataelement_15_z

Data model label MM_MISMATCH_REPAIR_GE

Level in data model `REQUIRED`

XSD name Mismatch repair gene expression

XSD type xs:string

List of permitted values in XSD

```
”LOSS_OF_EXPRESSION”
”EXPRESSION”
”NOT_DONE”
```

Type in data model

```
LIST_OF_VALUES [expression; loss of expression; not done]
```

XSD parent form Form - form_z8_ver-z7

XSD description

Mismatch repair gene expression – IHC array for different genes (common for 3). Expression of MLH1, MSHz, PMSz and MSH6

Data model description

Mismatch repair gene expression – IHC array for different genes (common for 3). Expression of MLH1, MSHz, PMSz and MSH6

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

4.7 Dataelement_16_3 - MM_RISK_SITUATION_HNPCC

XSD label Dataelement_16_3

Data model label MM_RISK_SITUATION_HNPCC

Level in data model `OPTIONAL`

XSD name Risk situation (only HNPCC)

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
YES_NO []  ((true|false|yes|no|f|t))
```

XSD parent form Form - form_z8_ver-z7

XSD description
Risk situation (only HNPCC), Amsterdam criteria

Data model description
Risk situation (only HNPCC), Amsterdam criteria

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

4.8 Dataelement_20_3 - MM_KRAS_MUTATION_KRAS_EX2

XSD label Dataelement_z0_3

Data model label MM_KRAS_MUTATION_KRAS_EXz

Level in data model `REQUIRED`

XSD name KRAS exon z (codons 1z or 13)

XSD type xs:string

List of permitted values in XSD

```
”Not mutated”
”Mutated”
”Not done”
```

Type in data model

```
LIST_OF_VALUES [Mutated; Not mutated; Not done]
```

XSD parent form Form - form_z8_ver-z7

XSD description
KRAS exon z (codons 1z or 13) mutation status

Data model description
KRAS exon z (codons 1z or 13) mutation status

4.9 Dataelement_21_5 - MM_KRAS_MUTATION_KRAS_EX3

XSD label Dataelement_z1_5

Data model label MM_KRAS_MUTATION_KRAS_EX3

Level in data model REQUIRED

XSD name KRAS exon 3 (codons 59 or 61)

XSD type xs:string

List of permitted values in XSD

”Not mutated”
”Mutated”
”Not done”

Type in data model

LIST_OF_VALUES [Mutated; Not mutated; Not done]

XSD parent form Form - form_z8_ver-z7

XSD description
KRAS exon 3 (codons 59 or 61) mutation status

Data model description
KRAS exon 3 (codons 59 or 61) mutation status

4.10 Dataelement_22_4 - MM_KRAS_MUTATION_KRAS_EX4

XSD label Dataelement_zz_4

Data model label MM_KRAS_MUTATION_KRAS_EX4

Level in data model `REQUIRED`

XSD name KRAS exon 4 (codons 117 or 146) mutation status

XSD type xs:string

List of permitted values in XSD

```
”Not mutated”
”Mutated”
”Not done”
```

Type in data model

```
LIST_OF_VALUES [Mutated; Not mutated; Not done]
```

XSD parent form Form - form_z8_ver-z7

XSD description
KRAS exon 4 (codons 117 or 146)

Data model description
KRAS exon 4 (codons 117 or 146)

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

4.11 Dataelement_23_5 - MM_KRAS_MUTATION_NRAS_EX2

XSD label Dataelement_z3_5

Data model label MM_KRAS_MUTATION_NRAS_EXz

Level in data model `REQUIRED`

XSD name NRAS exon z (codons 1z or 13)

XSD type xs:string

List of permitted values in XSD

```
”Not mutated”
”Mutated”
”Not done”
```

Type in data model

```
LIST_OF_VALUES [Mutated; Not mutated; Not done]
```

XSD parent form Form - form_z8_ver-z7

XSD description

NRAS exon z (codons 1z or 13) mutation status

Data model description

NRAS exon z (codons 1z or 13) mutation status

4.12 Dataelement_24_4 - MM_KRAS_MUTATION_NRAS_EX3

XSD label Dataelement_z4_4

Data model label MM_KRAS_MUTATION_NRAS_EX3

Level in data model `REQUIRED`

XSD name NRAS exon 3 (codons 59 or 61)

XSD type xs:string

List of permitted values in XSD

```
”Not mutated”
”Mutated”
”Not done”
```

Type in data model

```
LIST_OF_VALUES [Mutated; Not mutated; Not done]
```

XSD parent form Form - form_z8_ver-z7

XSD description
NRAS exon 3 (codons 59 or 61) mutation status

Data model description
NRAS exon 3 (codons 59 or 61) mutation status

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

4.13 Dataelement_25_3 - MM_KRAS_MUTATION_NRAS_EX4

XSD label Dataelement_z5_3

Data model label MM_KRAS_MUTATION_NRAS_EX4

Level in data model `REQUIRED`

XSD name NRAS exon 4 (cdons 117 or 146)

XSD type xs:string

List of permitted values in XSD

```
”Not mutated”
”Mutated”
”Not done”
```

Type in data model

```
LIST_OF_VALUES [Mutated; Not mutated; Not done]
```

XSD parent form Form - form_z8_ver-z7

XSD description
NRAS exon 4 (codons 117 or 146) mutation status

Data model description
NRAS exon 4 (codons 117 or 146) mutation status

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

4.14 Dataelement_2_2 - CLINICAL_STUDY_PARTICIPANT

XSD label Dataelement_z_z

Data model label CLINICAL_STUDY_PARTICIPANT

Level in data model OPTIONAL

XSD name Participation in clinical study

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
YES_NO []  ((true|false|yes|no|f|t))
```

XSD parent form Form - form_z8_ver-z7

XSD description
Participation in clinical study

Data model description
Participation in clinical study

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

4.15 Dataelement_30_3 - DIAG_MRI_DONE

XSD label Dataelement_30_3

Data model label DIAG_MRI_DONE

Level in data model `REQUIRED`

XSD name MRI

XSD type xs:string

List of permitted values in XSD

```
”MRI - Unknown”
”MRI - Done, data not available”
”MRI - Done, data available”
”MRI - Not done”
```

Type in data model

```
LIST_OF_VALUES [Done, data available; Done, data not available; Not


   * done; Unknown]
```

XSD parent form Form - form_z8_ver-z7

XSD description

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.

Data model description

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.

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

4.16 Dataelement_31_3 - DIAG_CT_DONE

XSD label Dataelement_31_3

Data model label DIAG_CT_DONE

Level in data model `REQUIRED`

XSD name CT

XSD type xs:string

List of permitted values in XSD

```
”CT - Unknown”
”CT - Done, data not available”
”CT - Done, data available”
”CT- Not done”
```

Type in data model

```
LIST_OF_VALUES [Done, data available; Done, data not available; Not

   * done; Unknown]
```

XSD parent form Form - form_z8_ver-z7

XSD description

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.

Data model description

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.

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

4.17 Dataelement_33_1 - THERAPY_RESPONSE

XSD label Dataelement_33_1

Data model label THERAPY_RESPONSE

Level in data model `REQUIRED`

XSD name Specific response

XSD type xs:string

List of permitted values in XSD

```
”Specific response - Complete response”
”Specific response - Partial response”
”Specific response - Stable disease”
”Specific response - Progressive disease”
```

Type in data model

```
LIST_OF_VALUES [Complete response; Partial response; Progressive


   * disease; Stable disease]
```

XSD parent form Form4 - form_31_ver-z

XSD description

Response to therapy - Specific response

Data model description

Response to therapy - Specific response

4.18 Dataelement_34_1 - THERAPY_RESPONSE_TIMESTAMP_RELATIVE

XSD label Dataelement_34_1

Data model label THERAPY_RESPONSE_TIMESTAMP_RELATIVE

Level in data model `REQUIRED`

XSD name Date response was obtained in weeks since initial diagnosis

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER []  (0<=x)
```

XSD parent form Form4 - form_31_ver-z

XSD description

Date response was obtained in weeks since initial diagnosis

Data model description

Date response was obtained in weeks since initial diagnosis

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

4.19 Dataelement_35_3 - TARGETED_THERAPY_START_RELATIVE

XSD label Dataelement_35_3

Data model label TARGETED_THERAPY_START_RELATIVE

Level in data model `REQUIRED`

XSD name Date of start of targeted therapy

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER []  (0<=x)
```

XSD parent form Form6 - form_30_ver-3

XSD description

Targeted therapy - Date of start (weeks since initial diagnosis)

Data model description

Targeted therapy - Date of start (weeks since initial diagnosis)

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

4.20 Dataelement_36_1 - TARGETED_THERAPY_END_RELATIVE

XSD label Dataelement_36_1

Data model label TARGETED_THERAPY_END_RELATIVE

Level in data model `OPTIONAL`

XSD name Date of end of targeted therapy

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER []  (0<=x)
```

XSD parent form Form6 - form_30_ver-3

XSD description
Targeted therapy - Date of end (weeks since initial diagnosis)

Data model description
Targeted therapy - Date of end (weeks since initial diagnosis)

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

4.21 Dataelement_3_1 - AGE_AT_PRIMARY_DIAGNOSIS

XSD label Dataelement_3_1

Data model label AGE_AT_PRIMARY_DIAGNOSIS

Level in data model `REQUIRED`

XSD name Age at diagnosis (rounded to years)

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [a]  (0<=x)
```

XSD parent form Form - form_z8_ver-z7

XSD description

Age at initial histopathological diagnosis (biopsy or surgical specimen of the primary tumor) rounded to years.

Data model description

Age at initial histopathological diagnosis (biopsy or surgical specimen of the primary tumor) rounded to years.

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

4.22 Dataelement_49_1 - SURGERY_TYPE

XSD label Dataelement_49_1

Data model label SURGERY_TYPE

Level in data model `REQUIRED`

XSD name Surgery type

XSD type xs:string

List of permitted values in XSD

```
”Other”
”Endo-rectal tumor resection”
”Abdomino-perineal resection”
”Anterior resection of rectum”
”Low anteroir colon resection”
”Pan-procto colectomy”
”Total colectomy”
”Sigmoid colectomy”
”Transverse colectomy”
”Left hemicolectomy”
”Right hemicolectomy”
```

Type in data model

```
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]
```

XSD parent form Form - form_3z_ver-8

XSD description

Surgery type

Data model description

Surgery type

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

4.23 Dataelement_4_3 - TIME_OF_RECURRENCE_RELATIVE

XSD label Dataelement_4_3

Data model label TIME_OF_RECURRENCE_RELATIVE

Level in data model `OPTIONAL`

XSD name Time of recurrence (metastasis diagnosis)

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [week]  (0<=x)
```

XSD parent form Form - form_z8_ver-z7

XSD description

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

Data model description

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

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

4.24 Dataelement_51_3 - DATE_DIAGNOSIS

XSD label Dataelement_51_3

Data model label DATE_DIAGNOSIS

Level in data model `OPTIONAL`

XSD name Date of diagnosis

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
DATE []  (ISO_8601_WITH_DAYS)
```

XSD parent form Form - form_z8_ver-z7

XSD description

Date at which colon cancer was diagnosed for the first time. Histopathological diagnosis by biopsy or surgery qualifies as primary diagnosis

Data model description

Date at which colon cancer was diagnosed for the first time. Histopathological diagnosis by biopsy or surgery qualifies as primary diagnosis

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

4.25 Dataelement_53_3 - WHO_GRADE_VERSION

XSD label Dataelement_53_3

Data model label WHO_GRADE_VERSION

Level in data model `REQUIRED`

XSD name WHO version

XSD type xs:string

List of permitted values in XSD

```
”Not known”
”1st edition”
”2nd edition”
”3rd edition”
”4th edition”
```

Type in data model

```
LIST_OF_VALUES [1st ed. (1979-1990); 2nd ed. (1991-2000); 3rd ed.

   * (2001-2010); 4th ed. (used since 2011); Edition not known]
```

XSD parent form Formz - form_34_ver-zz

XSD description

The version of the WHO classification system used.Version years:4th ed. (used since z011),3rd ed. (z001-z010),znd ed. (1991-z000),1st ed. (1979-1990)

Data model description

The version of the WHO classification system used

4.26 Dataelement_54_2 - SAMPLE_MATERIAL_TYPE

XSD label Dataelement_54_z

Data model label SAMPLE_MATERIAL_TYPE

Level in data model `REQUIRED`

XSD name Material type

XSD type xs:string

List of permitted values in XSD

```
”Other”
”Healthy colon tissue”
”Tumor”
```

Type in data model

```
LIST_OF_VALUES [Healthy colon tissue; Tumor tissue; Other]
```

XSD parent form Form1 - form_35_ver-6

XSD description
Type of specimen

Data model description
Type of specimen

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

4.27 Dataelement_55_2 - SAMPLE_PRESERVATION_MODE

XSD label Dataelement_55_z

Data model label SAMPLE_PRESERVATION_MODE

Level in data model `REQUIRED`

XSD name Preservation mode

XSD type xs:string

List of permitted values in XSD

```
”Other”
”Cryopreservation”
”FFPE”
```

Type in data model

```
LIST_OF_VALUES [Cryopreservation; FFPE; Other]
```

XSD parent form Form1 - form_35_ver-6

XSD description
The preservation mode for the specimen

Data model description
The preservation mode for the specimen

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

4.28 Dataelement_56_2 - SAMPLE_ID

XSD label Dataelement_56_z

Data model label SAMPLE_ID

Level in data model `REQUIRED`

XSD name Sample ID

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
TEXT [] ()
```

XSD parent form Form1 - form_35_ver-6

XSD description
An identifier, unique within the biobank

Data model description
An identifier, unique within the biobank

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

4.29 Dataelement_57_3 - DIGITAL_IMAGING_AVAILABILITY

XSD label Dataelement_57_3

Data model label DIGITAL_IMAGING_AVAILABILITY

Level in data model `OPTIONAL`

XSD name Availability digital imaging

XSD type xs:string

List of permitted values in XSD

```
”No”
”Can be generated”
”Readily available”
```

Type in data model

```
LIST_OF_VALUES [Can be generated; No; Readily available]
```

XSD parent form Formz - form_34_ver-zz

XSD description

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

Data model description

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 <0.1z5um/pixel (this is more accurate description of 40x).

4.30 Dataelement_58_2 - DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY

XSD label Dataelement_58_z

Data model label DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY

Level in data model `OPTIONAL`

XSD name Availability invasion front digital imaging

XSD type xs:string

List of permitted values in XSD

```
”Invasion front not included”
”Readily available”
”Can be generated”
”No”
```

Type in data model

```
LIST_OF_VALUES [Can be generated; Invasion front not included; No;

   * Readily available]
```

XSD parent form Formz - form_34_ver-zz

XSD description

Do you have high-resolution digital imaging (corresponding to magnification 40x) containing invasion front from the histopatology?

Data model description

Do you have high-resolution digital imaging (corresponding to magnification 40x) containing invasion front from the histopatology?

4.31 Dataelement_59_5 - PHARMACOTHERAPY_SCHEME

XSD label Dataelement_59_5

Data model label PHARMACOTHERAPY_SCHEME

Level in data model REQUIRED

XSD name Scheme of pharmacotherapy

XSD type xs:string

List of permitted values in XSD

```
”Other”
”Scheme of pharmacotherapy - 5-FU 325-350 mg/m2 + LV 20 mg/m2i.v.

    * bolus, day1-5, weeks 1 and 5”
”Scheme of pharmacotherapy - 5-FU 400 mg/m2 + 100 mg i.v. bolus, d

    ↪ 1,2, 11,12,21,22”
”Scheme of pharmacotherapy - 5-FU 225 mg/m2 i.v. continuous infusion,

    * 5 days per week”
”Scheme of pharmacotherapy - 5-FU 1000 mg/m2 i.v. continuous infusion,

    *  day 1-5, weeks 1 and 5”
”Scheme of pharmacotherapy - Capecitabine 800-825 mg/m2 bid po, day 1
    * -5, together with radiation or continuously untill end of

    * radiation”
”Scheme of pharmacotherapy - UFT (300-340mg/m2/day) and LV (22.5-90 mg
    * /day) po continuously, 5(-7) days per week, together with

    * radiotherapy”
”Scheme of pharmacotherapy - 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”
```

Type in data model

```
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/
```

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

```
↪ 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]
```

XSD parent form Form3 - form_33_ver-10

XSD description

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

Data model description

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

4.32 Dataelement_5_2 - VITAL_STATUS

XSD label Dataelement_5_z

Data model label VITAL_STATUS

Level in data model `REQUIRED`

XSD name Vital status

XSD type xs:string

List of permitted values in XSD

```
”ALIVE”
”DEATH_COLON_CANCER”
”DEATH_OTHER”
”DEATH_UNKNOWN_REASON”
”UNKNOWN”
```

Type in data model
`LIST_OF_VALUES [death due to colon cancer; death due to other reasons;`

- `death for unknown reasons; person is still alive; unknown]`

XSD parent form Form - form_z8_ver-z7

XSD description
Vital status

Data model description
Vital status

4.33 Dataelement_61_5 - DIAG_LIVER_IMAGING_DONE

XSD label Dataelement_61_5

Data model label DIAG_LIVER_IMAGING_DONE

Level in data model `REQUIRED`

XSD name Liver imaging

XSD type xs:string

List of permitted values in XSD

```
”Liver imaging - Unknown Not done”
”Liver imaging - Done, data available”
”Liver imaging - Done, data not available”
”Liver imaging - Unknown”
```

Type in data model

```
LIST_OF_VALUES [Done, data available; Done, data not available; Not


   * done; Unknown]
```

XSD parent form Form - form_z8_ver-z7

XSD description

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.

Data model description

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.

4.34 Dataelement_63_4 - DIAG_X_DONE

XSD label Dataelement_63_4

Data model label DIAG_X_DONE

Level in data model `REQUIRED`

XSD name Lung imaging

XSD type xs:string

List of permitted values in XSD

```
”Lung imaging - Not done”
”Lung imaging - Done, data available”
”Lung imaging - Done, data not available”
”Lung imaging - Unknown”
```

Type in data model

```
LIST_OF_VALUES [Done, data available; Done, data not available; Not


   * done; Unknown]
```

XSD parent form Form - form_z8_ver-z7

XSD description

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.

Data model description

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.

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

4.35 Dataelement_67_1 - SURGERY_TYPE_OTHER

XSD label Dataelement_67_1

Data model label SURGERY_TYPE_OTHER

Level in data model `OPTIONAL`

XSD name Other surgery type

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
TEXT [] ()
```

XSD parent form Form - form_3z_ver-8

XSD description
Surgery type, if not present on the list

Data model description
Surgery type, if not present on the list

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

4.36 Dataelement_68_2 - HIST_METASTASIS

XSD label Dataelement_68_z

Data model label HIST_METASTASIS

Level in data model `REQUIRED`

XSD name Localization of metastasis

XSD type Not defined.

List of permitted values in XSD

```
”Localization of metastasis - None”
”Localization of metastasis - Pulmonary”
”Localization of metastasis - Osseous”
”Localization of metastasis - Hepatic”
”Localization of metastasis - Brain”
”Localization of metastasis - Lymph nodes”
”Localization of metastasis - Bone marrow”
”Localization of metastasis - Pleura”
”Localization of metastasis - Peritoneum”
”Localization of metastasis - Adrenals”
”Localization of metastasis - Skin”
”Localization of metastasis - Others”
```

Type in data model

```
LIST_OF_VALUES [Adrenals; Bone marrow; Brain; Hepatic; Lymph nodes;
```

```
   * None; Osseous; Peritoneum; Pleura; Pulmonary; Skin; Others]
```

XSD parent form Formz - form_34_ver-zz

XSD description

Histopathology part - Localization of metastasis

Data model description

Histopathology part - Localization of metastasis. Multiple metastases can be added, each with its own location. This is intended for primary diagnosis only.

4.37 Dataelement_6_3 - VITAL_STATUS_TIMESTAMP

XSD label Dataelement_6_3

Data model label VITAL_STATUS_TIMESTAMP

Level in data model `if(VITAL_STATUS!=UNKNOWN){REQUIRED}else{OPTIONAL}`

XSD name Timestamp of last update of vital status

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
DATE []  (ISO_8601_WITH_DAYS)
```

XSD parent form Form - form_z8_ver-z7

XSD description

Timestamp of last update of vital status

Data model description

Timestamp of last update of vital status

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

4.38 Dataelement_70_2 - UICC_STAGE

XSD label Dataelement_70_z

Data model label UICC_STAGE

Level in data model REQUIRED

XSD name Stage

XSD type xs:string

List of permitted values in XSD

```
”Stage - IV”
”Stage - III”
”Stage - II”
”Stage - IVC”
”Stage - IIC”
”Stage - IVB”
”Stage - IVA”
”Stage - IIIC”
”Stage - IIIB”
”Stage - IIIA”
”Stage - IIB”
”Stage - II A”
”Stage - I”
”Stage - 0”
```

Type in data model

```
LIST_OF_VALUES [0; I; II; II A; IIB; IIC; III; IIIA; IIIB; IIIC; IV;


    * IVA; IVB; IVC]
```

XSD parent form Formz - form_34_ver-zz

XSD description

UICC Stage. The stages list is based on 8th edition, and backwards compatible with earlier editions.

Data model description

UICC Stage. The stages list is based on 8th edition, and backwards compatible with earlier editions.

4.39 Dataelement_71_1 - TNM_PRIMARY_TUMOR

XSD label Dataelement_71_1

Data model label TNM_PRIMARY_TUMOR

Level in data model `REQUIRED`

XSD name Primary Tumor

XSD type xs:string

List of permitted values in XSD

```
”Primary Tumor - T4”
”Primary Tumor - T4b”
”Primary Tumor - T4a”
”Primary Tumor - T3”
”Primary Tumor - T2”
”Primary Tumor - T1”
”Primary Tumor - Tis”
”Primary Tumor - T0”
”Primary Tumor - TX”
```

Type in data model

```
LIST_OF_VALUES [T0; T1; T2; T3; T4; T4a; T4b; Tis; TX]
```

XSD parent form Formz - form_34_ver-zz

XSD description

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

Data model description

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

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

4.40 Dataelement_73_3 - UICC_VERSION

XSD label Dataelement_73_3

Data model label UICC_VERSION

Level in data model `REQUIRED`

XSD name UICC version

XSD type xs:string

List of permitted values in XSD

```
”8th edition”
”Not known”
”7th edition”
”6th edition”
”5th edition”
”4th edition or earlier”
```

Type in data model

```
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]
```

XSD parent form Formz - form_34_ver-zz

XSD description

The version of the UICC system under which the staging was done. Version years:8th edition (since z017),7th edition (used z010-z017),6th. ed (used z003-z009),5th. ed (used 1998-z00z),4th. ed (used before 1998)

Data model description

The version of the UICC system under which the staging was done

4.41 Dataelement_75_1 - TNM_DISTANT_METASTASIS

XSD label Dataelement_75_1

Data model label TNM_DISTANT_METASTASIS

Level in data model `REQUIRED`

XSD name Distant metastasis

XSD type xs:string

List of permitted values in XSD

```
”Distant metastasis - MX”
”Distant metastasis - M1c”
”Distant metastasis - M0”
”Distant metastasis - M1”
”Distant metastasis - M1a”
”Distant metastasis - M1b”
```

Type in data model

```
LIST_OF_VALUES [M0; M1; M1a; M1b; M1c; MX]
```

XSD parent form Formz - form_34_ver-zz

XSD description

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

Data model description

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

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

4.42 Dataelement_77_1 - TNM_REGIONAL_LYMPH_NODES

XSD label Dataelement_77_1

Data model label TNM_REGIONAL_LYMPH_NODES

Level in data model `REQUIRED`

XSD name Regional lymph nodes

XSD type xs:string

List of permitted values in XSD

```
”Regional lymph nodes - N3”
”Regional lymph nodes - NX”
”Regional lymph nodes - N0”
”Regional lymph nodes - N1”
”Regional lymph nodes - N1a”
”Regional lymph nodes - N1b”
”Regional lymph nodes - N1c”
”Regional lymph nodes - N2”
”Regional lymph nodes - N2a”
”Regional lymph nodes - N2b”
```

Type in data model

```
LIST_OF_VALUES [N0; N1; N1a; N1b; N1c; N2; N2a; N2b; N3; NX]
```

XSD parent form Formz - form_34_ver-zz

XSD description

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

Data model description

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

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

4.43 Dataelement_7_2 - OVERALL_SURVIVAL_STATUS

XSD label Dataelement_7_z

Data model label OVERALL_SURVIVAL_STATUS

Level in data model `REQUIRED`

XSD name Overall survival status

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [week]  (0<=x)
```

XSD parent form Form - form_z8_ver-z7

XSD description

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

Data model description

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

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

4.44 Dataelement_81_3 - PHARMACOTHERAPY_SCHEME_DESCRIPTION

XSD label Dataelement_81_3

Data model label PHARMACOTHERAPY_SCHEME_DESCRIPTION

Level in data model `if(PHARMACOTHERAPY_SCHEME==Other){REQUIRED}else{OPTIONAL}`

XSD name Other pharmacotherapy scheme

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
TEXT [] ()
```

XSD parent form Form3 - form_33_ver-10

XSD description

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

Data model description

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

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

4.45 Dataelement_82_1 - BIOLOGICAL_MATERIAL_FROM_RECURRENCE_AVAILABLE

XSD label Dataelement_8z_1

Data model label BIOLOGICAL_MATERIAL_FROM_RECURRENCE_AVAILABLE

Level in data model `OPTIONAL`

XSD name Biological material from recurrence available

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
YES_NO []  ((true|false|yes|no|f|t))
```

XSD parent form Formz - form_34_ver-zz

XSD description

Biological material from recurrence available

Data model description

Biological material from recurrence available

4.46 Dataelement_83_1 - WHO_GRADE

XSD label Dataelement_83_1

Data model label WHO_GRADE

Level in data model `REQUIRED`

XSD name Grade

XSD type xs:string

List of permitted values in XSD

```
”WHO Grading - Grade - G1”
”WHO Grading - Grade - G2”
”WHO Grading - Grade - G3”
”WHO Grading - Grade - G4”
”WHO Grading - Grade - GX”
```

Type in data model

```
LIST_OF_VALUES [G1; G2; G3; G4; GX]
```

XSD parent form Formz - form_34_ver-zz

XSD description

Grade. For Sweden “medium high” shall map to G3, and “low medium” shall map to Gz. This has to be documented in the provenance information

Data model description

Grade. For Sweden “medium high” shall map to G3, and “low medium” shall map to Gz. This has to be documented in the provenance information

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

4.47 Dataelement_85_1 - SEX

XSD label Dataelement_85_1

Data model label SEX

Level in data model `REQUIRED`

XSD name Biological sex

XSD type xs:string

List of permitted values in XSD

```
”other”
”female”
”male”
```

Type in data model

```
LIST_OF_VALUES [female; male; other]
```

XSD parent form Form - form_z8_ver-z7

XSD description
Biological sex of the person, defined by chromosomes.

Data model description
Biological sex of the person, defined by chromosomes.

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

4.48 Dataelement_87_1 - BRAF_PIC3CA_HER_MUTATION_STATUS

XSD label Dataelement_87_1

Data model label BRAF_PIC3CA_HER_MUTATION_STATUS

Level in data model `OPTIONAL`

XSD name BRAF, PIC3CA, HERz mutation status

XSD type xs:string

List of permitted values in XSD

```
”BRAF, PIC3CA, HER2 mutation status - Partial information available”
”BRAF, PIC3CA, HER2 mutation status - not mutated”
”BRAF, PIC3CA, HER2 mutation status - mutated”
”BRAF, PIC3CA, HER2 mutation status - not done”
```

Type in data model
`LIST_OF_VALUES [Mutated; Not mutated; Partial information available;`

* `Not done]`

XSD parent form Form - form_z8_ver-z7

XSD description
BRAF, PIC3CA, HERz mutation status. If only 1 or z of the three mutation analyses have been done, the “Partial information available” value shall be selected

Data model description
BRAF, PIC3CA, HERz mutation status. If only 1 or z of the three mutation analyses have been done, the “Partial information available” value shall be selected

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

4.49 Dataelement_88_1 - DIAG_COLONOSCOPY

XSD label Dataelement_88_1

Data model label DIAG_COLONOSCOPY

Level in data model `REQUIRED`

XSD name Colonoscopy

XSD type xs:string

List of permitted values in XSD

```
”Colonoscopy diagnostic exam- Unknown”
”Colonoscopy diagnostic exam- Not done”
”Colonoscopy diagnostic exam - Negative”
”Colonoscopy diagnostic exam - Positive”
```

Type in data model

```
LIST_OF_VALUES [Negative; Positive; Not done; Unknown]
```

XSD parent form Form - form_z8_ver-z7

XSD description

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.

Data model description

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.

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

4.50 Dataelement_89_3 - YEAR_OF_SAMPLE_COLLECTION

XSD label Dataelement_89_3

Data model label YEAR_OF_SAMPLE_COLLECTION

Level in data model `REQUIRED`

XSD name Year of sample collection

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [years]  (0<=x)
```

XSD parent form Form1 - form_35_ver-6

XSD description

Calendar year in which the sample was collected.(YYYY)

Data model description

Calender year in which the sample was collected.

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

4.51 Dataelement_8_3 - SURGERY_START_RELATIVE

XSD label Dataelement_8_3

Data model label SURGERY_START_RELATIVE

Level in data model `REQUIRED`

XSD name Time difference between initial diagnosis and surgery

XSD type xs:string

List of permitted values in XSD Not defined.

Type in data model

```
NATURAL_NUMBER [week]  (0<=x)
```

XSD parent form Form - form_3z_ver-8

XSD description

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.

Data model description

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.

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

4.52 Dataelement_91_1 - HIST_MORPHOLOGY

XSD label Dataelement_91_1

Data model label HIST_MORPHOLOGY

Level in data model `REQUIRED`

XSD name Morphology

XSD type xs:string

List of permitted values in XSD

```
”Signet ring cell carcinoma”
”Cribriform comedo-type adenocarcinoma”
”Adenocarcinoma”
”Mucinous carcinoma”
”Signet-ring cell carcinoma”
”Medullary carcinoma”
”High-grade neuroendocrine carcinoma”
”Large cell neuroendocrine carcinoma”
”small cell neuroendocrine carcinoma”
”Squamous cell carcinoma”
”Adeonsquamous carcinoma”
”Micropapillary carcinoma”
”Serrated adenocarcinoma”
”Spindle cell carcinoma”
”Mixed adenoneuroendocrine carcinoma”
”Undifferentiated carcinoma”
”Other”
```

Type in data model

```
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]
```

XSD parent form Formz - form_34_ver-zz

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

XSD description

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

Data model description

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

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

4.53 Dataelement_92_1 - HIST_LOCALIZATION

XSD label Dataelement_9z_1

Data model label HIST_LOCALIZATION

Level in data model `REQUIRED`

XSD name Localization of primary tumor

XSD type xs:string

List of permitted values in XSD

```
”Localization of primary tumor - C20”
”Localization of primary tumor - C19”
”Localization of primary tumor - C18.7”
”Localization of primary tumor - C18.6”
”Localization of primary tumor - C18.5”
”Localization of primary tumor - C18.4”
”Localization of primary tumor - C18.3”
”Localization of primary tumor - C18.2”
”Localization of primary tumor - C18.1”
”Localization of primary tumor - C18.0”
```

Type in data model

```
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]
```

XSD parent form Formz - form_34_ver-zz

XSD description

Histopathology part - Localization of primary tumor

Data model description

Histopathology part - Localization of primary tumor

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

4.54 Dataelement_93_1 - SURGERY_LOCATION

XSD label Dataelement_93_1

Data model label SURGERY_LOCATION

Level in data model `REQUIRED`

XSD name Location of the tumor

XSD type xs:string

List of permitted values in XSD

```
”Location of  tumor - C18.0”
”Location of  tumor - C18.1”
”Location of  tumor - C18.2”
”Location of  tumor - C18.3”
”Location of  tumor - C18.4”
”Location of  tumor - C18.5”
”Location of  tumor - C18.6”
”Location of  tumor - C18.7”
”Location of  tumor - C19”
”Location of  tumor - C19.9”
”Location of  tumor - C20”
”Location of  tumor - C20.9”
```

Type in data model

```
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]
```

XSD parent form Form - form_3z_ver-8

XSD description

Location of the tumor

Data model description

Location of the tumor

4.55 Dataelement_9_2 - SURGERY_RADICALITY

XSD label Dataelement_9_z

Data model label SURGERY_RADICALITY

Level in data model `REQUIRED`

XSD name Surgery radicality

XSD type xs:string

List of permitted values in XSD

```
”R2”
”R1”
”R0”
”RX”
```

Type in data model

```
LIST_OF_VALUES [R0; R1; R2; RX]
```

XSD parent form Form - form_3z_ver-8

XSD description
Whether the surgery removed the entire tumor.

Data model description
Whether the surgery removed the entire tumor.

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

C. Colorectal Cancer Collection Participation Letter

European-wide collection of colorectal cancer cases --- Follow up and information letter for the biobanks

1. Background

BBMRI---ERIC,[1] 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,** where the access to European biobanks is piloted through a colorectal cancer cohort use case (gathering of existing colon cancer data sets/samples 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, samples for future research use. The data sets are gathered, anonymized, and made available centrally for the research community to query and identify their specific research questions in colorectal cancer (access provision explained later on this document). The access to the centrally stored data is therefore provided to any researcher who whishes to define their research question based on this data, however, the access to more granular data (additional data that becomes of interest), or later on the samples linked to any of the data, will be accessible only through the official access procedures of the particular biobanks. **The purpose of the cohort is 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
- enable the connection between the European biobank's information with the coded clinical IT systems
- provide the research community an opportunity to define and tackle pertinent research questions in the field of colon cancer research
- demonstrate the benefits of operational distributed Research Infrastructure to advance high---quality research and innovation

The exact research questions this data can serve are yet to be defined, but steps are taken to open up a new chapter in pan---European research on cancer.

2. Why we contact you?

As the first step to establish the colorectal cancer cohort, all BBMRI---ERIC Member States mapped the biobanks in their national networks that 1) have colorectal cancer samples, 2) would be interested in providing the samples and associated data for research use, and further 3) defined the types of samples they have on colorectal cancer. As a second step, **BBMRI---ERIC now contacts all National Nodes and the biobanks who have already expressed their interest to become part of this European wide cohort gathering. With this letter, we provide the biobanks detailed information of the requested clinical information important for their potential participation.**

[1] BBMRI---ERIC Biobanking and Biomolecular resources Research Infrastructure---European Research Infrastructure Consortium

Horizon 2020

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

3. What do we gather precisely and how?

In total 10,000 cases of colorectal cancer cases shall be gathered in selected European biobanks that have FFPE (formalin---fixed paraffin embedded, minimum requirement) or fresh frozen tissue samples of surgical material available with associated clinical data (see attached the REQUIRED and OPTIONAL data items; Appendix I and II). 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 IT---portal offered by BBMRI---ERIC (a tool for manual entry of data). The rest of the 7,000 cases are collected semi---automatically by using data---mining and/or 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. The data will be gathered coded (pseudonymized) and will be stored centrally at CNR---ITB (Istituto di Tecnologie Biomediche Consiglio Nazionale delle Ricerche, Italy) in a protected environment suitable for storing pseudonymized/ anonymized data under Italian legislation for future use.

The 3,000 cases – manual collection:

For the 3,000 colorectal cancer cases that are collected manually from the biobanks through an IT---portal, BBMRI---ERIC will reimburse the participating biobanks with 150 EUR per case (reimbursement only for complete clinical data sets and potential provision of the tissue material for research projects if/when requested). The IT---portal for the collection of the cases has been built and will be released soon.

The 7,000 cases – semi---automated collection:

For the remaining 7,000 colorectal cancer cases, data mining tools will be deployed and customized to fit the local specifics of some selected biobanks that already have advanced IT systems and personnel. This will be done in conjunction with the Common Service IT of BBMRI---ERIC. The biobanks participating may receive up to equivalent of 6 person months financial support for interfacing the provided tools to their biobank and/or hospital information systems. This will be however decided in case to case basis.

The REQUIRED data set:

Colorectal cancer experts have defined a comprehensive data set (see the Appendix I and II) 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. The required data items will be collected as **coded** and before providing them for any generic searching, the data will be **anonymized** using state of the art anonymization techniques (the data sets provided by the different biobanks will also serve for benchmarking anonymization techniques).

Horizon 2020

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

What is in it for your biobank?

- The collection of colorectal cancer cohort will be a joint European endeavor providing biobanks ***recognition, visibility, new users, data*** and ***high quality research results.***
- BBMRI-ERIC community will aim to illustrate and promote the ***prominent role of biobanks in driving the medical research and related IT-solutions forward.***
- Limited monetary support is available for those biobanks that able to supply manually entered data and for some biobanks interfacing the text mining tools with their IT systems (see manual/semi-automated collection).
- The biobanks that are able to apply and fully utilize the text-mining tools will have ***an opportunity to use the tools,*** if applicable, ***for a broad spectrum of diseases in the future.***
- Being part of the unique European wide cohort on colorectal cancer will attract attention within the research community and steps are taken to open up ***a new chapter in pan-European research on cancer.***

4. What happens after the data is gathered? – The access

BBMRI---ERIC operates ***as a facilitator in the provision of access*** *to the gathered anonymized REQUIRED and OPTIONAL colorectal cancer datasets, while* ***the local biobanks stay in control of granting access to samples or any additional data*** *based on their existing access policies, proper peer review and ethical assessment of research proposals*. Hence, the access is realized in two levels 1) access to the REQUIRED and OPTIONAL data sets only (BBMRI---ERIC controls) 2) and access to the samples and or additional data (biobank controls).

The REQUIRED data sets are stored centrally at CNR---ITB in a protected environment suitable for storing pseudonymized/anonymized data under Italian legislation for future use. The samples and/or any additional data are stored in their original storages (e.g. local biobanks). To access the REQUIRED and OPTIONAL data sets through BBMRI---ERIC, the accessor needs to follow the criteria and procedure of BBMRI---ERIC.

5. The collection of data is realized in phases

The aim of this letter is to contact the National Nodes and those biobanks who have already expressed their interest to become part of the colorectal cancer cohort collection and invite them to partake **the pilot collection**.

Pilot phase: The partaking is based on biobanks' capacity, capability, and willingness to contribute to the collection of colorectal cancer data both methods: manual collection and semi---automated collection will be used.

The purpose of the pilot is to:

- collect data through the manual IT---portal in order to train the text/data mining tools and to validate the results;
- adapt the validated text mining tools to be operational in each participating biobank;
- collect data using validated text mining tools;
- understand and overcome the hurdles of the collection process;
- prepare for the main run of the data collection

Collection phase: After the pilot is finished, we will perform the main run of the data collection.

6. Timeline

March 2017

- Contacting the biobanks who have expressed their interest to participate the colorectal cancer data collection.
- Obtaining direct connection with the relevant people in the biobanks (manager/IT personnel /researcher/ medical expert).
- Selecting the biobanks suitable for the pilot phase.

April---May 2017

- Preparing for the data collection (e.g. links to registries/ ethical approvals and other necessary steps)

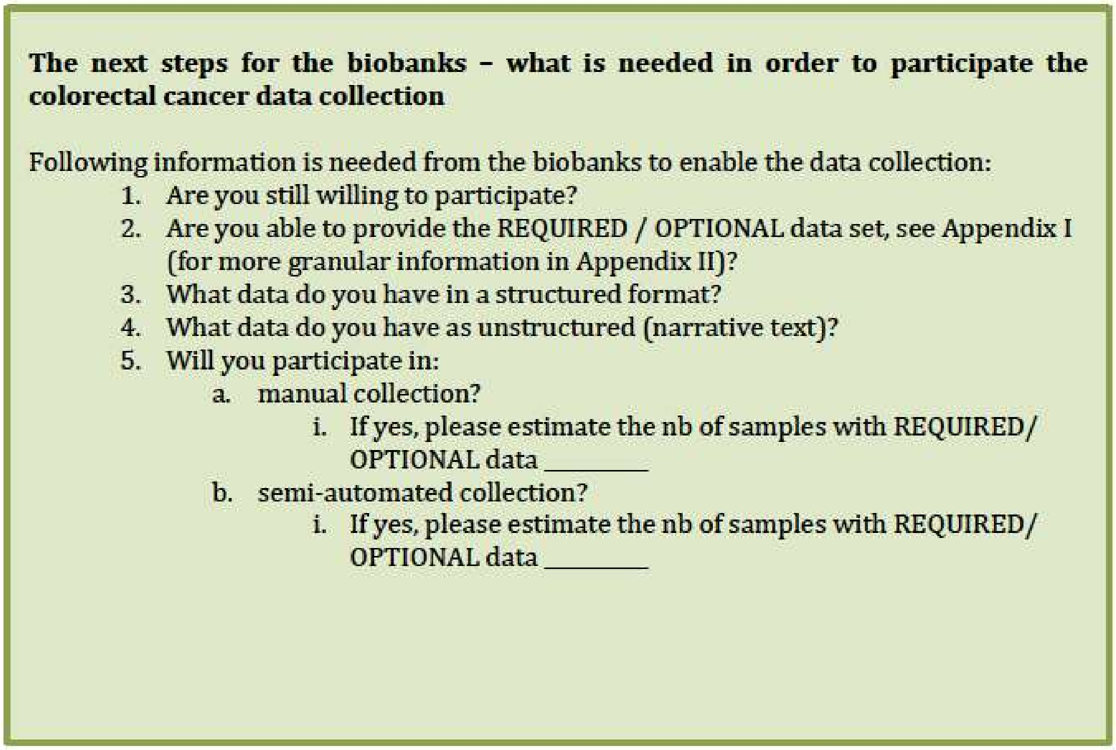
June---September 2017

- Adjusting the text/data mining tools for biobanks – if required
- Running the data collection
- Preparing for the main run of the data collection

October 2017---March 2018

- Main run of data collection

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



The next steps for the biobanks – what is needed in order to participate the colorectal cancer data collection

Following information is needed from the biobanks to enable the data collection:

1. Are you still willing to participate?
2. Are you able to provide the REQUIRED / OPTIONAL data set, see Appendix I (for more granular information in Appendix II)?
3. What data do you have in a structured format?
4. What data do you have as unstructured (narrative text)?
5. Will you participate in:
   a. manual collection?
      i. If yes, please estimate the nb of samples with REQUIRED/ OPTIONAL data ________
   b. semi-automated collection?
      i. If yes, please estimate the nb of samples with REQUIRED/ OPTIONAL data ________

Send your response by 31 March 2017 to:

marialuisa.lavitrano@unimib.it
michael.hummel@charite.de
cc to:
outi.tornwall@bbmri---eric.eu

Contact:

IT specific questions:	petr.holub@bbmri-eric.eu
Colorectal cancer specific questions:	michael.hummel@charite.de marialuisa.lavitrano@unimib.it

General questions about the collection and the pilot:
outi.tornwall@bbmri---eric.eu
marialuisa.lavitrano@unimib.it
michael.hummel@charite.de

Appendix I – Required dataset for colorectal cancer collection – overview
Appendix II – Detailed description of the required data set (excel file)

Appendix I – Required dataset for colorectal cancer collection – overview

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

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 (C18.1 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 MTA/DTA.

Definition of the data items

Defined variables:

- Sex:
- Participation in clinical study
- Age at primary diagnosis:
- Time of recurrence (metastasis)
- Vital status and survival information
  - Timestamp of last update of vital status
  - Overall survival status
- Surgery: aggregate object
  - Time difference between initial diagnosis and surgery:
  - Surgery radicality:
  - Type of surgery:
- Pharmacotherapy:
  - REQUIRED if occurred
  - Date of start:
  - Scheme of pharmacotherapy:
- Targeted therapy:
  - REQUIRED if occurred
  - Date of start:
  - Date of end:
- Radiation therapy:

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

  - 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
  - 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
  - morphology
  - Localization
  - Metastasis
  - High resolution (40x) digital image preferably from invasion front
- Diagnostic exam
  - Colonoscopy
  - Array of diagnostic methods (liver imaging, lung imaging, MRI, CT)

D. CRC Biobanks

Country	Coordinator	Biobank	Institution/City	Responsible person	email(s)	Grouping	PARTICIPATION TO PILOT -status	Data structured	Data unstructured	Method of participation	Estimated nb of samples
Italy	Marialuisa Lavitrano	Centro Risorse Biologiche Istituto Clinico Humanitas	Istituto Clinico Humanitas-IRCCS, Milano	Daniela Pistillo	*daniela.pistillo@cancercenter.humanitas.it*	***1***	*Yes to participation, are able to provide the data, but the IT management need more info*	*all*	*no*	***manual+semi-automated***	500
Austria	Kurt Zatloukal	Biobank Graz	Medical Universtity of Graz, Graz	Berthold Huppertz	*berthold.huppertz@medunigraz.at*	***1***	*Yes to participation, have required data*	*all*	*all*	***manual+ semi-automated***	1400
Italy	Marialuisa Lavitrano	CRO BIOBANK	CRO Centro di Riferimento Oncologico, NCI, IRSSC Aviano	Paolo de Paoli, contact: Vincento Canzonieri	*dirscienti@cro.it, vcanzonieri@cro.it*	***2***	*Yes to participation, have all of the required data*	*All*	*no*	*manual*	180
Italy	Marialuisa Lavitrano	ARC-Net Applied Research on Cancer Network	University Hospital Trust of Verona, Verona	Aldo Scarpa / Rita T. Lawlor	*aldo.scarpa@univr.it , ritateresa.lawlor@univr.it*	***2***	*Yes to participation, have required data*	*newer data yes*	*older mostly + histopathology, molecular biology*	*manual*	600
Belgium	Annelies Debucquoy	Bordet Tumor Bank	Institut Jules Bordet	Denis Larsimont	*denis.larsimont@bordet.be*	***2***	*yes to participation, able to provide the data, (fresh frozen tissues available)*	*some*	*mostly*	*manual*	500
Belgium	Annelies Debucquoy	Tumorbank@uza VZW	Universitair Ziekenhuis Antwerpen	CEO: Dhr Johnny Van der Straeten;	*Sofie.Goethals@uza.be, PRIMARY CONTACT: stephanie.gofflot@chu.ulg.ac.be*	***2***	*Yes to participation, are able to provide most of the data (??). Questions about ethical approval, how would that work*	*localization meta and primary tumour, morphology, pTNM, UICC version, mutation statuses of molecular markers, sex, data of diagnosis, sample ID, material type, preservation mode*	*mostly*	*manual*	80
Belgium	Annelies Debucquoy	Biothèque Hospitalo Universitaire de Liège	CHU de Liège	Philippe Delvenne	*P.Delvenne@ulg.ac.be PRIMARY CONTACT stephanie.gofflot@chu.ulg.ac.be*	***2***	*yes to participation, able to provide the data, (fresh frozen tissues available)*	*Histopathology and samples data*	*surgery and theraphy*	*manual*	10
Finland	Anu Jalanko	Helsingin Biobankki	Helsinki ja Uusimaa Hospital District	Kimmo Pitkanen	*kimmo.pitkanen@hus.fi*	***2***	*Yes to participation, are able to provide the requested/optional data.*	*pathology (snomed, pTMN, IHC, RAS pathway analysis, lab values, comorbidities ICD10, medication)*	*the rest of the data*	*manual*	500
Italy	Marialuisa Lavitrano	Genetic Biobank of Siena (GBS)	Azienda Ospedaliera Universitaria Senese and University of Siena, Siena	Alessandra Renieri	*alessandra.renieri@unisi.it*	***2***	*Yes to participation, they have required data*	*NO*	*all*	*manual*	200
France	Michael Hisbergues	CRB-CH	Perpignan	A. GARRET-GIRAUDON	*aline.garret@ch-perpignan.fr*	***2***	*Yes to participation, are able to provide the requested/optional data.*	*NO*	*all*	*manual*	40
UK	Philip Quinlan	The Northern Ireland Biobank	Queen's University Belfast, Northern Ireland, UK	Jacqueline James (contact: Priscilla Clark)	*nibiobank@qub.ac.uk*	***2***	*Yes to participation, have both required and optional data.*	*all*	*no*	*manual*	300
Italy	Marialuisa Lavitrano	SAVE	Istituto per lo Studio e la Prevenzione Oncologica, Firenze	Francesca Carozzi	*f.carozzi@ispo.toscana.it*	***2***	*Yes to participation, have required data*	*clinical data*	*Protocol collection instructions*	*manual*	4
Sweden	**Mats Hansson**	Uppsala Biobank	Uppsala Clinical Research Center	Anna Beskow / Tobias Sjöblom	*tobias.sjoblom@igp.uu.se (per-henrik.edqvist@igp.uu.se)*	***2***	*Yes to participation, have required data (KRAS mutation and MMR may be hard to get)*	*mostly*	*only few parameters (e.g. date of tretment)*	*manual*	800
Cyprus	**Kyriacos Kyriacou**	CING Biobank	The Cyprus Institute of Neurology and Genetics	Kyriacos Kyriacou	*kyriacos@cing.ac.cy*	***2***	*Yes to participation, have required data*	*clinical data*	*histopathology, molecular data*	*manual*	200
Italy	Marialuisa Lavitrano	Centro Risorse Biologiche CRB-USMI	IRCCS AOU San Martino – IST, Genova	Barbara Parodi	*barbara.parodi@hsanmartino.it*	***3***	*Yes to participation, are able to provide the data*	*sample data*	*partially clinical data*	*manual (250)+semi-automated (200)*	250
Italy	Marialuisa Lavitrano	INT Biobank	Fondazione IRCCS Istituto Nazionale dei Tumori, Milano	Mariagrazia Daidone	*mariagrazia.daidone@istitutotumori.mi.it*		*Yes to participate*			*manual+semi-automated*	500
Check Republic	Dalibor Valík	Masaryk Memorial Cancer Institute	Masaryk Memorial Cancer Institute		*nenutil@mou.cz*	***3***	*Yes to participation, have required data*	*Histopathology+TNM, WHO classification, Patient data, pharmacotheraphy, sample, vital status*	*Diagnostic exam, Histopathology UICC, molecul markers, response to theraphy, targetted theraphy*	*manual+semiautomated*	300
France	Michael Hisbergues	Tumorothèque France-Comté	Besançon	S. VALMARY-DEGANO	*sdeganovalmary@chu-besancon.fr*	***3***	*Yes to participation, have required data*	*sample, patient and histopathology data*	*diagnostic exam, pharmaco theraphy, response, surgery, vital status,*	*manual and semiautomated*	300
Belgium	Annelies Debucquoy	BruTuS, Bruxelles	CHU Brugmann	Ruth Duttmann	*ruth.duttmann@chu-brugmann.be (PRIMARY CONTACT:Maxime.LORENT@chu-brugmann.be)*	***unclear***	*Yes to participation, have required data*	*protocols*	*mostly*	*No information available*	10
Belgium	Annelies Debucquoy	Biothèque Institut Roi Albert II	Cliniques universitaires Saint-Luc	Etienne Marbaix	*etienne.marbaix@uclouvain.be*	***unclear***	*Intersted, but not able to commit to transferring any data without ehtical/scientific approvals.*	*histopathology*	*mostly*	*manual*	10
Belgium	Annelies Debucquoy	Biobank of CHU Dinant Godinne	CHU Dinant Godinne|UCL Namur	Carlos GRAUX	*carlos.graux@uclouvain.be (PRIMARY CONTACT: Fabienne.George@uclouvain.be)*	***unclear***	*Yes to participation, are able to provide the data (but NO digital imaging)*	*some*	*mostly*	*manual*	
Germany	Michael Hummel	Interdisciplinary Bank of Biomaterials and Data Würzburg (ibdw)	University Hospital and University of Würzburg	Roland Jahns	*Jahns_R@ukw.de*	***unclear***	*Yes*	*No information available*	*No information available*	*manual*	1000
Germany	Michael Hummel	UCT Biobank Frankfurt	University Cancer Center Frankfurt	(Hans Michael Kvasnicka, the head of bb) Daniel Brucker	*Daniel.Brucker@kgu.de, Brandts@em.uni-frankfurt.de*	***unclear***	*Yes*	*No information available*	*No information available*	*manual+semi-automated (50+50)*	100
Germany	Michael Hummel	Biobank of National Center for Tumor Diseases	University Hospital Heidelberg, Institute of Pathology	Esther Herpel	*esther.herpel@med.uni-heidelberg.de, romy.kirsten@nct-heidelberg.de*	***unclear***	*Yes*	*No information available*	*No information available*	*manual*	300

Grouping:

1 = Data structured, manual+semi-automated collecson - no data mining

2 = Data unstructured, manual collecson

3 = Data unstructured, manual+semi-automated collecson - data mining

Pilot Groups	Sample count
1	2080
2	3234
3	1350
unclear	1420
Total	8084