

ADOPT BBMRI-ERIC GRANT AGREEMENT NO 676550

DELIVERABLE REPORT

Deliverable no	D2.7
Deliverable Title	Patient-related data for colorectal cancer samples
Contractual delivery month	M42 (March 2019)
Responsible Partner	BBMRI.it
Author(s)	Petr Holub, Marialuisa Lavitrano, Outi Törnwall, Rumyana Proynova, Florian Stampe, Maxmilian Ataian, Irene Schlünder, Anna-Liisa Kuslap, Caitlin Ahern, Erik Steinfelder, Kurt Zatloukal, Rudolf Nenutil, Luciano Milanesi, Michael Hummel

PATIENT-RELATED DATA FOR COLORECTAL CANCER SAMPLES

Executive Summary

This Deliverable summarizes achievements and results obtained during the collection of Colorectal Cancer Cohort (CRC-Cohort), a Europe-wide effort to collect a cohort of at least 10 000 colorectal cancer patients within the ADOPT project. The collected data set is made available for better visibility and accessibility of the contributing BBMRI-ERIC biobanks. The total number of collected cases reached 10 480, provided by 25 biobanks from 12 European countries, providing good geographical coverage of Europe. This deliverable provides basic information on the methods used to collect the cohort, on IT tools developed for the collection process, as well as on aggregate statistical information on the collected cohort.

The effort began in March 2016 with the definition of the data set to be collected. The main data collection period ran from January 2018 to March 2019. The delivery date was extended with Amendment #2 from M36 to M42 due to the delay in collecting the colon cancer cohort.

The process of collecting the CRC-Cohort comprised several steps. First, BBMRI-ERIC biobanks were queried regarding interest in participation. Second, a working group of medical and IT experts convened to define the common data model (D2.4), towing the line between clearly defined data structures while providing enough information for meaningful medical research.

The third step, composing the Data Protection Policy, raised many issues and challenges but was ultimately successful. (D2.1 contains a summary of the result, and the final policy is published as D2.3 Annex III.)

In the fourth step, the policies were distributed to member biobanks so that they could determine their ability to provide samples to the CRC-Cohort.



The subsequent steps were the Implementation of the resulting data model in the common Metadata Repository (MDR), Implementation of central data collections system called Colorectal Cancer Data Collection system (CCDC), Design and implementation of data harmonization tools to support the conversion process from common tabular files, Design and implementation of data quality checks in collaboration between expert pathologists and IT experts, and finally, the Data quality improvement cycle.

The deliverable then provides an in-depth discussion of the developed reimbursement model, which turned out to be significantly more complex than originally anticipated. Since almost all biobanks qualified for semi-automated extraction mode, meaning that a substantial amount of data was already available in a structured form, it was decided to use a backup plan already prepared in the ADOPT project proposal; the reimbursement model was adjusted to reflect these changes. The resulting model is a linear combination of UNIMIB and BBMRI-ERIC funding sources: each biobank gets a proportion of the cases reimbursed as manually delivered and part as delivered using automated processing; due to the two funding sources with different fixed reimbursement rates the model is relatively complex.

Results of the deliverable are numerous and extend beyond collecting over the targeted 10,000 colorectal cancer datasets. The CRC-Cohort Data Protection Policy was released in October 2017 and used in the dataset collection process. It supports biobanks by offering guidelines on a range of topics such as data access and quality assurance.

Furthermore, the data model defined by the interdisciplinary expert working group has been implemented in the MDR, where it is available via an API for access by other components of the CRC-Cohort ecosystem of IT tools. The central Colorectal Cancer Data Collection system was implemented based on open-source software extended adequately to support the CRC-Cohort and to feature both graphical user interface (via web) or an API for programmatic upload of the data into the system. The web-based user interface was anticipated for the small biobanks contributing data manually only. The system has been deployed on BBMRI-ERIC production IT infrastructure managed within Common Service IT.

In addition, a system of data quality checks was developed as a central service running on the central database, with 70 different data checks covering consistency and suspicious data. Further checks for missing data uploaded by the biobanks were also created.

Exceeding the initial target, ADOPT managed to collect 10 480 cases of colorectal cancer from 25 biobanks in 12 European countries, thus covering an unprecedented area of patients, from the UK to Finland to Cyprus. The datasets have been anonymised, aggregated and analysed in terms of survival rates, availability of molecular markers, and therapy events.

The conclusion admits that the demands far exceeded initial expectations for several reasons: the significant discrepancy in availability of structured in-depth data in different European Countries, in organizational requirements, as well as in availability of IT expertise to manipulate the data at source. Technical and organizational measures to ensure data security and protect privacy of the persons contributing their data to CRC-Cohort were discussed in-depth and the agreement on data transfer was reached in compliance with the General Data Protection Regulation (GDPR), with the IMI Code of Practice on Secondary Use of Medical Data in Scientific Research Projects and taking into consideration the differences in regulatory and ethical issues within the different European Countries.

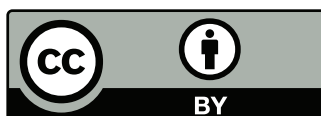


Lessons learned from the deliverable are chronicled for future reference. These include the greater-than-expected amount of time needed both for initiating the data collection process and for approval of contracts. Furthermore, the importance of robust legal support, working in harmony with IT, cannot be understated in order to create contracts that allow for the sharing of data while upholding national legal requirements. The need for data quality checks was apparent and will need to be addressed in future projects in which data remains in source repositories and thus cannot be checked by the federated research infrastructure. With federated querying systems such as the BBMRI-ERIC Locator, the biobanks should be contractually bound to update their own data; still, it will be difficult for a central entity to ensure quality for each participating biobank under this model. Biobanks also need to be warned that they will need to allocate significant resources to continuous data quality improvement, not only to primary data collection.

All lessons learned will be taken into consideration when improving the organization of BBMRI-ERIC biobanks and when designing and developing IT tools to make the biological material and data compliant with FAIR and FAIR-Health principles. The resulting CRC-Cohort of 10,480 datasets is being made available for researchers in compliance with the access modes defined in the CRC-Cohort Data Protection Policy (see ADOPT Deliverable D2.3 Appendix III).



COPYRIGHT NOTICE



This work by Parties of the ADOPT BBMRI-ERIC Consortium is licensed under a Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 676550.

DOCUMENT LOG

Issue	Date	Comment	Author
D2.7rev1	2019-03-20	Initial version	Petr Holub
D2.7rev2	2019-09-18	Revised version	Petr Holub, Caitlin Ahern



Glossary

API Application Programming Interface

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

BBMRI-ERIC Directory A service provided by BBMRI-ERIC to enable basic findability of biobanks and their collections of samples/data. <https://directory.bbmri-eric.eu/>

BBMRI-ERIC Locator A service co-developed by BBMRI-ERIC and German Biobank Alliance to enable advanced findability of biobanks and their collections based on sample-level and donor-level data. <https://http://search.germanbiobanknode.de/>

CCDC A system developed as a part of ADOPT project to collect and host Colorectal Cancer Cohort (CRC-Cohort) data

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

Common Service IT Common Service on Information Technologies (IT)

CRC-Cohort Colorectal Cancer Cohort, the subject of this document

MDR Repository of metadata developed in OSSE Project¹ based on ISO/IEC 11179 standard [3].

National/Organisational Node A National Node or an Organisational Node as defined in the Statutes of BBMRI-ERIC.

partner biobank Biobanks partnering with BBMRI-ERIC via National/Organisational Node, as defined in the Statutes of BBMRI-ERIC, Article 1, §10 [1, Annex 1]

¹ <https://www.osse-register.de/en/>



Contents

Background	7
1. Methods	8
1.1. Process of developing Colorectal Cancer Cohort (CRC-Cohort)	8
1.2. Updated reimbursement model	13
2. Results	16
2.1. Colorectal Cancer Cohort (CRC-Cohort) Data Protection Policy	16
2.2. Implementation of data model in Metadata Repository (MDR)	16
2.3. Colorectal Cancer Data Collection system (CCDC) database and data collection application	17
2.4. Data quality analyzer	18
2.5. Collected data set	18
3. Discussion and Conclusions	26
3.1. Lessons learned.	26
Bibliography	29
A. XML Format for CRC-Cohort data import into Colorectal Cancer Data Collection system (CCDC)	30



Background

This Deliverable summarizes achievements and results obtained during the collection of CRC-Cohort, a Europe-wide effort to collect a cohort of at least 10 000 colorectal cancer patients within the ADOPT project. The collected data set is made available for better visibility and accessibility of the contributing BBMRI-ERIC biobanks.

This effort of forming a cohort with existing colorectal cancer cases with detailed pathological and clinical data and available tissue samples is to demonstrate the feasibility of large scale collaboration within BBMRI-ERIC and generate a yet unprecedented resource for medical research. 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 delivery date was extended with Amendment #2 to M42 due to the delay in collecting the colon cancer cohort. Originally it was planned in M36.



1. Methods

1.1. Process of developing Colorectal Cancer Cohort (CRC-Cohort)

The whole process of building the CRC-Cohort comprised a number of steps as follows:

1. **Advertisement of the effort to BBMRI-ERIC biobanks** and preliminary inquiry of their interest to participate.

This was an initial step to define the initial set of biobanks with which further communication was done; this group was not meant to be closed, as some biobanks were expected to drop out due to finding. At this point only specification was the description of the cohort design as written in the ADOPT proposal.

2. **Definition of the common data model** created by an interdisciplinary working group of the following medical and IT experts. The data model focused on unambiguous definition of the data structure so that it can be implemented in IT systems, and on defining which parts of the data model are required to obtain data set meaningful for medical research.

The working group consisted of the following experts:

- *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;
- *IT experts:* Petr Holub, Frank Ückert, Diogo Alexandre, Ondřej Vojtíšek.

and the resulting model has been presented in the ADOPT Deliverable D2.4 (not part of this deliverable).

The process involved a series of teleconferences of the whole group, with specific follow up breakout meetings with particular subgroups (e.g., oncologists describing possible treatments and outcomes of treatments, molecular biologists defining the most relevant genetic variations). The design process was completely consensus driven: there were no items that could not be resolved during the initial design phase.

3. **Development of CRC-Cohort Data Protection Policy** (Petr Holub, Irene Schlünder, Kurt Zatloukal, Outi Törnwall, Marialuisa Lavitrano, Michael Hummel) – see also Section 2.1 for summary of the result. The proposed policy was consulted with the biobanks that indicated their interest in participating in the CRC-Cohort. CRC-Data-Protection Policy has been published as ADOPT D2.3 Appendix III.

Designing Data Protection Policy was a complex process which required number of important decisions and related documents:

- **Decision on the role** of BBMRI-ERIC in the CRC-Cohort with respect to the data protection: data processor vs. data controller.



Decision of the role was not clear upfront, as collection of the data set itself could have been achieved technically using either of those roles. The reasons why BBMRI-ERIC decided to opt for the data controller eventually were several: (a) it is the BBMRI-ERIC that sets the purpose of the data collection process and not the biobanks (data controllers); (b) CRC-Cohort becoming a permanent asset in order to improve visibility of the biobanks, while being a data processor on behalf of the controller is a time-limited activity; and (c) internationally accepted templates of contracts are commonly available for the data controller role and well understood (e.g., coming from BioMedBridges project for the domain of life sciences research infrastructures), while data processor agreements are typically very specific for a particular purpose of data processing.

Becoming a data controller has, however, created a problem for Finnish biobanks, as the current Finnish legislation forbids transfer of data controllership outside of Finland (for this reason Finnish data is missing from other long-term international resources such as dbSNP). A specific solution was then developed to allow Finnish biobanks to participate in CRC-Cohort: BBMRI-ERIC acts as a data processor to verify availability of the data and check data quality. After that the data is deleted and it is responsibility of the participating Finnish biobanks to make the data findable using the same means as BBMRI-ERIC implements for the rest of the CRC-Cohort.

- **Development of Data Provider Agreement template** defining the relation between BBMRI-ERIC and each biobank contributing to the CRC-Cohort as shown in Figure 1. With the exception of Finnish biobanks as described above, the Data Provider Agreements are contracts defining BBMRI-ERIC as data controller for the CRC-Cohort. The initial template was used as a proposal to the biobanks and was subject to negotiation. Individual modifications were typically introduced per request of each biobank, while striving to minimize deviations in order to have consistent contracting situation across the whole CRC-Cohort.

While small wording changes were introduced as requested individually by the biobank, the only substantial modifications content-wise were as follows:

- Italian biobanks needed explicit statement that for any access requests with potential commercial aspect, they need to be involved in the data release procedure so that their ethical review board can review the particular request (this was anyway the case as defined in the access procedure below, but they needed to have it as an explicit clause in the main part of the contract).
 - Finnish biobanks contribute to the cohort based on the controller–processor relation and the contract is based on their standard contracting template modified to suit the purpose of CRC-Cohort. Finnish biobanks also had to file their standard project application form to be reviewed and approved and this application has been developed jointly by the BBMRI-ERIC and Finnish staff.
- **Development of Data Transfer Agreement template** designed to cover situation when somebody requests personal data from the CRC-Cohort directly via BBMRI-ERIC. This contract is a controller–processor contract. It is assumed, however, that primary purpose of the CRC-Cohort is to increase findability of the source biobanks and that for more advanced research purposes the requesters will need to access more detailed data set directly from the biobanks. The contracting would be done directly between the requester and the biobank in such a case



and the centrally collected data would server to simplify data harmonization for already collected parts of the requested data set. Hence the Data Transfer Agreement template is only meant for those less common cases where the CRC-Cohort is directly used.

- **Development of Service Contract** to provide an infrastructure to host and process the CRC-Cohort data set. This step is needed as BBMRI-ERIC uses capacity of its National/Organisational Nodes to operate IT infrastructure as a part of its Common Service IT.

During the duration of the ADOPT project, BBMRI-ERIC used IT infrastructure at CNR, Milan, Italy, as the infrastructure is provided by the same team also as a part of BBMRI-ERIC Common Service IT.

- **Development of access procedure** aligned with the Data Protection Policy as well as BBMRI-ERIC Access Policy. The main difference of the access procedure compared to the common procedure applied when accessing the data from the biobanks (where BBMRI-ERIC participated only in the initial negotiation phase and then the contracts are directly implemented between the requester and the biobank), the BBMRI-ERIC runs the complete access procedure and only asks the contributing biobanks about the specific part of the requested data set: biobanks have the right, within one month, to veto the release of the part of the requested data set that was contributed by the given biobank. This unusual “veto with timeout” mechanism still substantially improves the situation compared to other common access requests, where BBMRI-ERIC has no control over behavior of the biobanks and can’t provide any guarantees on their reaction times.
- **Development of findability mechanisms** aligned with the Data Protection Policy, based on querying aggregate data via BBMRI-ERIC Directory or more detailed searches by querying anonymized data set via BBMRI-ERIC Locator, as described in depth in the Section 3 of the Data Protection Policy. Both of these are expected to result in increased visibility of the contributing biobanks.

All of these documents except for the Service Contract eventually became part of the appendices of the Data Protection Policy document and were distributed to the biobanks for approval.

4. **Recruitment process of BBMRI-ERIC biobanks and obtaining feedback on the proposed data model,** given the heterogeneity of European health care and medical standards.

In this step all the main documents were available for biobanks to do in-depth analysis if they can participate in the CRC-Cohort and to provide more accurate estimates of the amount of cases they can contribute: namely the detailed initial data model and the Data Protection Policy including all the appendices described above.

It was necessary to give biobanks enough time (up to 2 months) to review and comment on the Policy, as it typically had to undergo review by an ethics review board they adhere to and by their institutional legal department. Obtained feedback was used to improve Policy.

The first version of the data model was distributed to the biobanks to pilot it on their available data, in order to obtain real-world feedback on the problems the process will face once the actual data



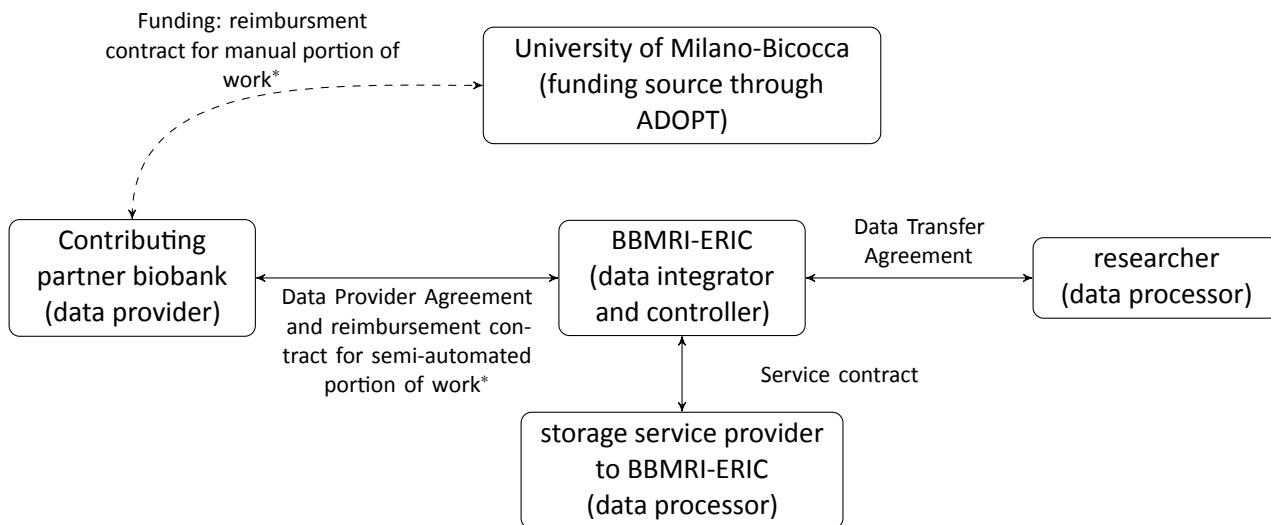


Figure 1: Contracting situation for the CRC-Cohort.

collection starts. All the feedback from the biobanks was collected, resulting in 57 comments as shown in the Appendix A of Deliverable D2.4 – **Data Model Review Log**. Each of those comments were processed in collaboration with the specific domain experts from the original design group; the review log document contains responses to all the requests, including information if it resulted in some modification of the model and explanation to the requester.

5. **Implementation of the resulting data model in the common Metadata Repository (MDR)** [3], where is in machine-readable form to be used by applications. See also Section 2.2 for the description of the result.

This is a common approach in the medical informatics domain that the data models are implemented in a dedicated MDR² in order to reuse those in multiple different systems and to be able to update the data model consistently in such a system. However, the practical implementation revealed a encounters one common design problem: the data model is often documented on the level of attributes of entities, while the relations of the entities are part of the design of the target system.³ This includes the fact that (at least simple) attribute constraints can be part of the MDR while relations on cardinalities of relations cannot be part of it. Hence the data model coming from MDR is incomplete and needs to be complemented with additional information in order to enable biobanks to contribute practically.

6. **Implementation of central data collections system called Colorectal Cancer Data Collection system (CCDC)**, including database, web-based user interface for manual contributions and API to allow programmatic imports of the data. See also Section 2.3 for the description of the result.

² See ISO/IEC 11179.

³ BBMRI-ERIC has observed similar behavior patterns also in the MIABIS, a community-driven standardization efforts in the biobanking domain, where experts from biobanks often focus on the attributes and only realize later how incomplete the interoperability is if the entities and their relations are not well-defined.



This process ran in parallel with the previous steps after designing the initial version of the data model, and the resulting system was later updated to the revised version of the model. This parallelism was necessary as even when basing the system on existing Samplify system, the implementation of the relatively complex data model, optimization of interfaces for manual data entry interface for both speed and data entry effectiveness, as well as creating the documentation for the whole system including API, required more than half a year of focused work.

It was necessary to work with the programmers and data managers to expect iterative changes to the data model and hence changes to the whole system, in order to avoid developers frustration on series of consecutive changes. This can be easily explained to the programmers in the terms of agile development models, yet the complexity of having users sometimes with conflicting needs (differing opinions of medical experts when updating the data model, conflicting needs of biobanks due to different national healthcare systems, etc.) is non-trivial to manage. It is also necessary to prototype real versions of the software for the biobanks to test, as often only after work with the real system the most useful feedback comes, as biobanks typically do not have advanced data modeling experts and need to work with real data on a real system (hence the iterations can be done on “pure theoretical” level of the data model).

7. **Design and implementation of data harmonization tools** to support the conversion process from common tabular files (Excel, CSV/TSV files, etc.). This is subject to ADOPT Deliverable D3.5 and [4].

Majority of the biobanks still unfortunately lack advanced IT systems and capabilities to do data transformation into the target formats and validate the results, once more complex data formats are required, such as deeply structured XML or JSON files. CRC-Cohort ended up with this type of files because of relatively rich data set including relations between entities such as treatments and responses to the treatments. Most of the data files used by the biobanks is still comma/tab-separated values (CSV/TSV files respectively) or Excel sheets. Furthermore it is necessary to map the values from their source data models to the data model defined in the common data set.

In order to support this process a dedicated toolset was developed [4], which achieves both the data mapping task and the task to convert the data to the target XML format needed for importing the data via API into the CCDC system. The system maintains a database of mappings, which are reusable in case that either new data arrives from the same biobank, or if starting with a new biobank coming from a similar national/regional context.

The tool was eventually operated mostly directly by the central IT team, as biobanks were only interested in delivering the data in their native formats. Consulting with the biobank was, however, needed in order to have correct interpretation and mappings of values.

8. **Design and implementation of data quality checks** in collaboration between expert pathologists and IT experts, as described more in-depth in Section 2.4 for the description of the checks.

The checks were run on the CCDC server automatically in a periodic manner and the results are sent back to the respective biobank after they contribute new or updated data. The reports are produced as Excel files since this format is the most approachable for the biobanks. This was typically followed up by consultations with the biobanks to interpret the data quality check report correctly.



It was quite surprising for the whole team how problematic the data from the biobanks were in some cases – typical issues were inconsistent survival information provided (e.g. treatments provided even after patient died), mismatching TNM values with UICC stage values (which is relatively simple check based on UICC standard), mismatch between location of tumor and type/location of surgery, and various problems related to outcomes of treatment (like starting a new chemotherapy after complete response to therapy, without indicating any recurrence). Some other data quality issues became obvious only after pooling larger amounts of data from multiple biobanks together, so that substantial differences in distributions became more visible (e.g., analyzing overall survival data with respect to the stage).

9. **Data quality improvement cycle**, where the centrally collected data set was updated based on data received from the biobanks based on the data quality reports. Most of the time the biobanks updated the complete data set and uploaded it as a new one, hence the old data was overwritten by the new data.

The last two steps have been repeated as until either all problems were fixed, or biobanks acknowledged that the detected problems are false positives or they are no longer able to improve quality of the data. Some biobanks complained that the iterative nature of this process posed a substantial burden for them and wanted to contractually cap the number of data quality check iterations. While this looks as a reasonable request initially, there is a two problems that it creates: (a) new upload of the data after fixing some problems triggers another positive problem report (despite during the implementation we tried to report as many problems at once as possible - sometimes at cost of reporting two or more different interdependent problems as separate reports that were fixed by correcting single data issue only); and (b) new data quality checks being developed even after the initial implementation phase which tried to be as complete as possible. A compromise was typically found with focus on fixing the most problematic data in the first rounds.

1.2. Updated reimbursement model

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 that have substantial part of the data already structured in machine readable formats. This included support provided to the biobanks in order to help them develop algorithmic transformations of structured data (see also step 7), as well as providing machine interface allowing batch upload of the data to the CCDC system (in step 6). Survey among the biobanks (step 4) revealed that most of the biobanks already have substantial parts of the data available in structured form, with 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. Therefore almost all biobanks qualified into semi-automated extraction mode.

Consistently with the ADOPT proposal, the reimbursement model was adjusted to reflect these changes. The *resources available for reimbursing manual work* of the biobank, which were part of University of Milano-Bicocca (UNIMIB) budget in WP2 (total of 450 000 € for 3000 cases at 150 € per case), were combined with the *budget for automated processing of records*, which was part of BBMRI-ERIC budget in



WP3 (total of 175 000 € for 7000 cases at 32 € per case). The resulting model is a linear combination of those two funding sources: each biobank gets proportion of the cases reimbursed as manually delivered and part as delivered using automated processing. The model is relatively complex and was necessary to combine the two funding sources that had different fixed reimbursement rates.

Parametrization of the model was further influenced by the decision of the Management Committee of BBMRI-ERIC and Director General, to pay 23.6 € *expedite bonus* to early contributors to the CRC-Cohort. This expedite bonus was capped to 1820 cases overall, with no single biobank being able have more than 300 cases reimbursed with this bonus; this approach ensures that at least $1820/300 = 6$ biobanks can obtain the expedite bonus funding. Eventually 9 biobanks obtained the bonus and the number of bonified cases ranged between 38 to 300 cases (see Figure 2).

Therefore the following model has been established based on combination of the two resources:

Mode	% cases reimbursed from UNIMIB	% cases reimbursed from BBMRI-ERIC	Total reimbursed cases	Average reimbursement per case [€]
<i>normal</i>	25	75	8000	61.5
<i>expedite delivery</i>	45	55	1820	85.1
<i>small biobanks</i>	100	0	180	150

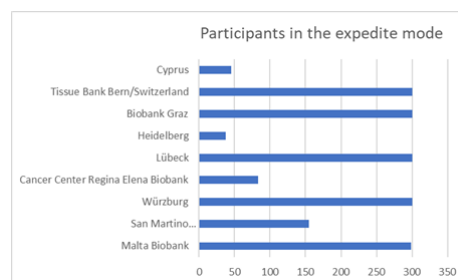


Figure 2: Reimbursement of biobanks in expedite mode. The graph shows how many cases were reimbursed in the expedite mode for each of the biobanks which managed to get the expedite delivery bonus.

For small biobanks, which were contributing fewer than 50 cases, the overhead of setting up semi-automated conversion was typically disproportionately large and hence the biobanks were reimbursed for manual processing of all the data (this is denoted as *small biobanks* row in the table). This eventually encompassed 4 biobanks contributing 10, 10, 20, and 40 cases respectively.

For the rest of the biobanks, the following algorithm was used. We use the following notation: n_{expedite} ...number of cases reimbursed with the expedite bonus of 23.6 €, while n_{normal} ...number of cases reimbursed at normal rate. Obviously $n_{\text{expedite}} + n_{\text{normal}}$ is the total number of cases for other than small biobanks.

$$\text{reimbursement} = \underbrace{[n_{\text{expedite}} * .45] * 150}_{\text{from UNIMIB}} + \underbrace{[n_{\text{expedite}} * .55] * 32}_{\text{from BBMRI-ERIC}} + \underbrace{[n_{\text{normal}} * .25] * 150}_{\text{from UNIMIB}} + \underbrace{[n_{\text{normal}} * .75] * 32}_{\text{from BBMRI-ERIC}}$$



The 23.6 € expedite bonus is achieved by increasing the proportion of cases reimbursed from UNIMIB, where the reimbursement per case is higher. By this it was possible to increase the average reimbursement case, while still having constant reimbursement rates for both sources of funding (UNIMIB and BBMRI-ERIC). There were minor deviations from the target average rates because of rounding of numbers of cases (the ceil functions), which were compensated by the BBMRI-ERIC so that the contractual reimbursement rate defined between BBMRI-ERIC and contributing biobank in the Data Provider Agreement was adhered to.

Example of reimbursement. Let's take the example of Biobank Graz for illustration purposes (first biobank in alphabetical order). They were contributing 1066 cases overall with 300 cases contributed with expedite bonus, hence this calculation results in the following:

- 135 cases being reimbursed from UNIMIB for expedite delivery at 150 € per case;
- 165 cases being reimbursed from BBMRI-ERIC for expedite delivery at 32 € per case;
- 191 cases being reimbursed from UNIMIB for normal delivery at 150 € per case;
- 575 cases being reimbursed from BBMRI-ERIC for normal delivery at 32 € per case.

Note that $135 + 165 = 300$ and $135 + 165 + 191 + 575 = 1066$. The average reimbursement rate for expedite cases is hence $(135 * 150 + 165 * 32)/300 = 85.1$ for expedite delivery and $(191 * 150 + 575 * 32)/(191 + 575) = 61.4229765$ for normal delivery (compared to 61.5 € in an ideal case).

Because of rounding in the normal mode, there is a rounding difference of $(575+191)*(61.5-61.4229765) = 59$ € (approx .08% of their total target reimbursement of 72 639 €), compensated by BBMRI-ERIC.



2. Results

The collection of 10 000 datasets related to colorectal cancer patients' samples from multiple biobanks is one of the major achievement of ADOPT. The patient-related datasets have been gathered, anonymized, and made available centrally for the research community to query and identify their specific research questions in colorectal cancer.

2.1. Colorectal Cancer Cohort (CRC-Cohort) Data Protection Policy

The data protection policy has been developed in order to allow the contributing biobanks to request approval by their governing bodies. This policy has been developed by Petr Holub, Irene Schlünder, Kurt Zatloukal, Outi Törnwall, Marialuisa Lavitrano, and Michael Hummel. It 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. Full version of the policy has been published as ADOPT D2.3 Appendix III.

The data protection policy gathers the objectives of the CRC cohort, its legal framework and basic organizational aspects. It describes the data collection and integration process, together with measures for quality checking and assurance.

In order to be a self-contained document to support biobanks in their participation in the CRC-Cohort, it covers a broad range of topics; the most important ones are:

- definition of purpose and scope of the project;
- description of data collection and storage process, including use of privacy enhancing technologies to ensure due data protection;
- data access regime;
- risk analysis;
- overview of the data model for collecting the data;
- data provider agreement template (between the contributing biobank and BBMRI-ERIC) and data transfer agreement template (a template for future contracts between BBMRI-ERIC and a researcher willing to do research with the CRC-Cohort data).

2.2. Implementation of data model in Metadata Repository (MDR)

The data model defined by the interdisciplinary expert working group has been implemented in the MDR [3, 6], where it is available via an API for access by other components of the CRC-Cohort ecosystem of IT tools. The MDR instance hosting the data model is publicly available at <https://mdr.osse-register.de/view.xhtml?namespace=ccdg>. Availability of the data model in a machine-readable structure is a prerequisite in order to have the whole system FAIR compliant in the future.

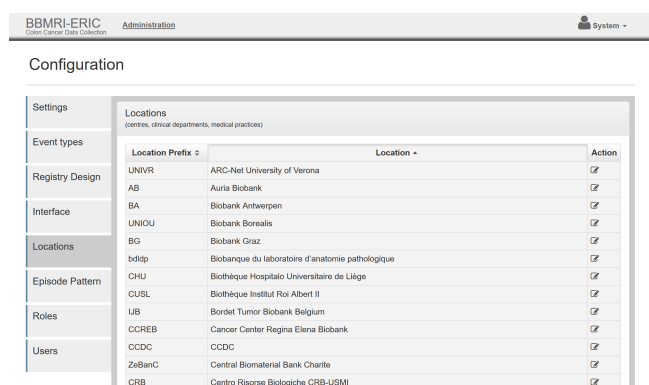
This work has been done by Rumyana Proynova.



2.3. Colorectal Cancer Data Collection system (CCDC) database and data collection application

The central CCDC system for collecting data was implemented based on open-source software coming from OSSE Project⁴ [5]. The software was extended adequately to support the CRC-Cohort and to feature both graphical user interface (via web) or an API for programmatic upload of the data into the system.

The web-based user interface, as shown in Figure 3, was anticipated for the small biobanks contributing data manually only. It was also possible to add or correct missing data via the web UI for the biobanks contributing the initial data via upload, but this method was discouraged for reasons of very complicated maintenance of this hybrid upload/manual data entry.



The screenshot shows the 'Configuration' page of the CCDC administrative interface. It features a sidebar with navigation options: Settings, Event types, Registry Design, Interface, Locations, Episode Pattern, Roles, and Users. The main content area is titled 'Locations (centres, clinical departments, medical practices)' and contains a table with columns for 'Location Prefix', 'Location', and 'Action'.

Location Prefix	Location	Action
UNIVR	ARC-Net University of Verona	<input type="checkbox"/>
AB	Auria Biobank	<input type="checkbox"/>
BA	Biobank Antwerpen	<input type="checkbox"/>
UNIOU	Biobank Borealis	<input type="checkbox"/>
BG	Biobank Graz	<input type="checkbox"/>
bdisp	Biobanque du laboratoire d'anatomie pathologique	<input type="checkbox"/>
CHU	Biothèque Hospitalo Universitaire de Liège	<input type="checkbox"/>
CUSL	Biothèque Institut Roi Albert II	<input type="checkbox"/>
IJB	Bordet Tumor Biobank Belgium	<input type="checkbox"/>
CCREB	Cancer Center Regina Elena Biobank	<input type="checkbox"/>
CCDC	CCDC	<input type="checkbox"/>
ZeBanc	Central Biomaterial Bank Charité	<input type="checkbox"/>
CRB	Centro Risorsa Biologica CRB-USMI	<input type="checkbox"/>

Figure 3: Administrative web interface of CCDC

The API uses REST with XML data payload. Two XSD-based data validation schema have been developed: one for full strict validation, which detects also any missing data required by the data mode, and one for partial validation, which does not take completeness into account and focuses on validating already provided data. The latter mode allows biobanks to upload incomplete data, obtain feedback from the data quality checks, and subsequently improve the data set handed over to BBMRI-ERIC; alternatively, the source biobanks can provide the missing data via the web-based UI.

The system has been deployed on BBMRI-ERIC production IT infrastructure managed within Common Service IT. For contributing biobankers it is available at <https://ccdc.bbmri-eric.eu/> using secure HTTPS protocol over public Internet. For administrators it is only accessible after two factor authentication only for any management operations (first factor being VPN authentication required for administrative access), the other being authentication using local accounts on the server via secure SSH protocol. The CCDC is hosted in the IT Infrastructure (server) of the CNR of Italy, which is linked to BBMRI-ERIC by long-term relationship defined as a part of Common Service IT. CNR guarantee the maintenance of the infrastructure as well as security and access to the patient-related data for colorectal cancer samples.

⁴ <https://www.osse-register.de/en/>



2.4. Data quality analyzer

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 has been developed as a central service running on the central database, with the following areas covered:

- *Checks for missing data* as the biobanks were allowed to provide incomplete data sets initially that should be completed by either editing the data via web UI, or providing updated data set for batch import. These checks analyze the data set based on items declared as REQUIRED in the data model.
- *Consistency checks* such as
 - negative durations of treatments;
 - mapping of pTNM values to UICC stage based on the UICC standard;
 - detector new series of treatments without indicating recurrence;
 - inconsistencies within surgical data (such as location of tumor and surgery type);
 - mismatch between histopathological and surgical data;
 - incomplete follow-up (such as patient has died of colorectal cancer, but the last response to treatment being complete response).
- *Suspicious data detector* including unlikely situations like extremely long survival of patients, big differences between survival last reported follow-up of the patient, very similar values of survival for big groups of patients (e.g., one biobank provided the same survival for all alive patients), or suspicious values in description of pharmacotherapy, combination of certain pTNM values with the UICC stage (e.g., pNX combined with defined stage, where it is not clear on what basis the stage was determined),

There are currently 70 different types of check in place, without counting the checks for missing data.

Primary authors of the data quality checks design were Petr Holub, Rudolf Nenutil, and Kurt Zatloukal, and the actual implementation was done by Florian Stampe and Petr Holub.

2.5. Collected data set

Total of 10 480 cases was collected, provided by 25 biobank from 12 European countries, providing good geographical coverage of the whole Europe (also see Figure 4). Per-biobank contribution was as follows:

Biobank	Cases
Auria Biobank	1,344
Central Biomaterial Bank Charite	1,226
Biobank Graz	1,066
Uppsala Biobank	1,017
Gewebe Biobank - Bern	871
Helsinki Biobank	653
ibdw Wuerzburg	646



Malta Biobank	533
INT Biobank	500
Interdisziplinäres Centrum für Biobanking-Lübeck	417
Humanitas Cancer Center	397
Nottingham Health Science Biobank	308
Masaryk Memorial Cancer Institute	300
Deutsches Krebsforschungszentrum	300
Cancer Center Regina Elena Biobank	218
TrentinoBioBank	218
Centro Risorse Biologiche CRB-USMI	155
Biobank Antwerpen	69
CRO BIOBANK	55
CING Biobank	50
Biobanque du laboratoire d'anatomie pathologique	50
Medical University of Gdansk	47
CHU UCL Namur	20
Biotheque Hospitalo Universitaire de Liege	10
CHU Brugmann	10

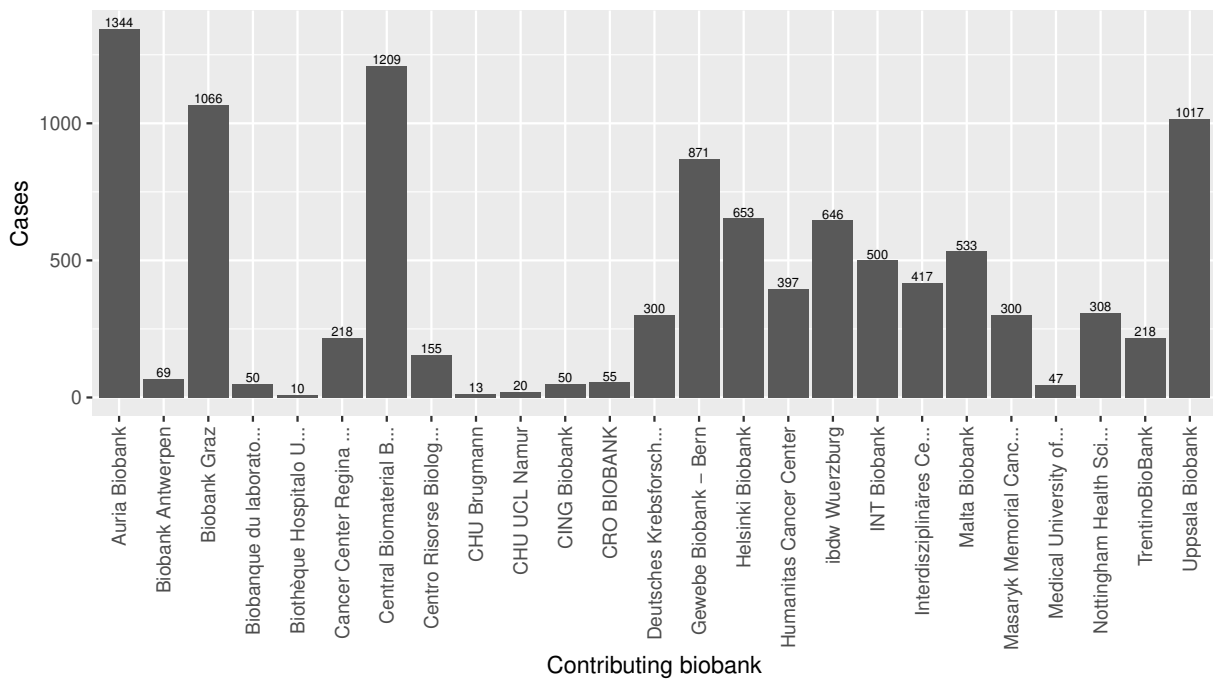


Figure 4: Contribution distribution per biobank.



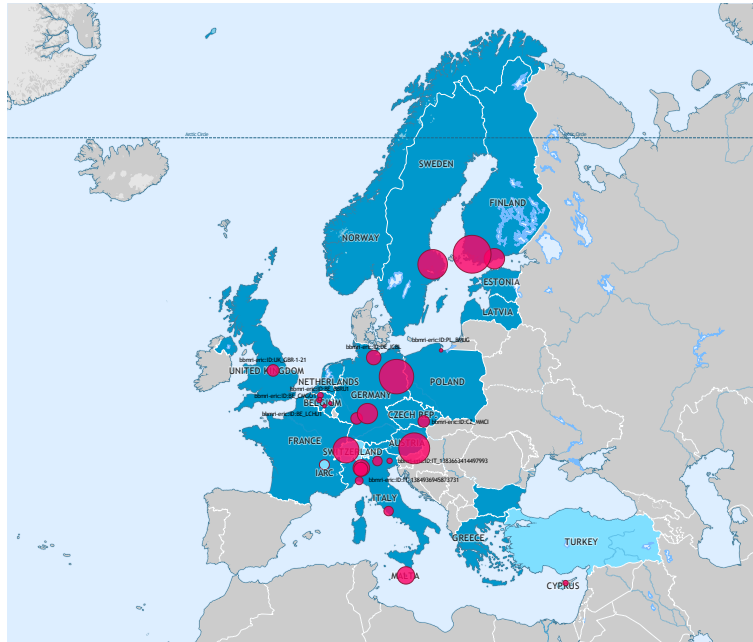


Figure 5: Geographical distribution of the contributing biobanks and the extent of their contribution. The size of the circle is proportionate to the number of contributed cases.

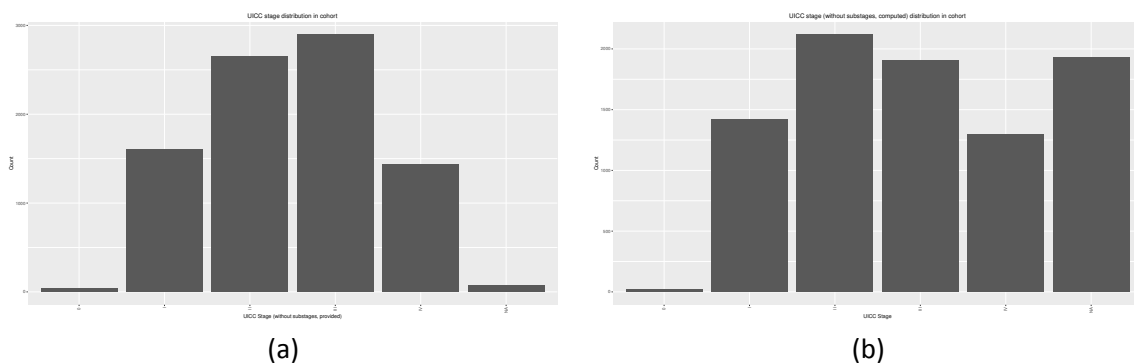


Figure 6: Distribution of available UICC stages (a) as provided by biobanks, (b) as computed from the pTNM values provided by biobanks.



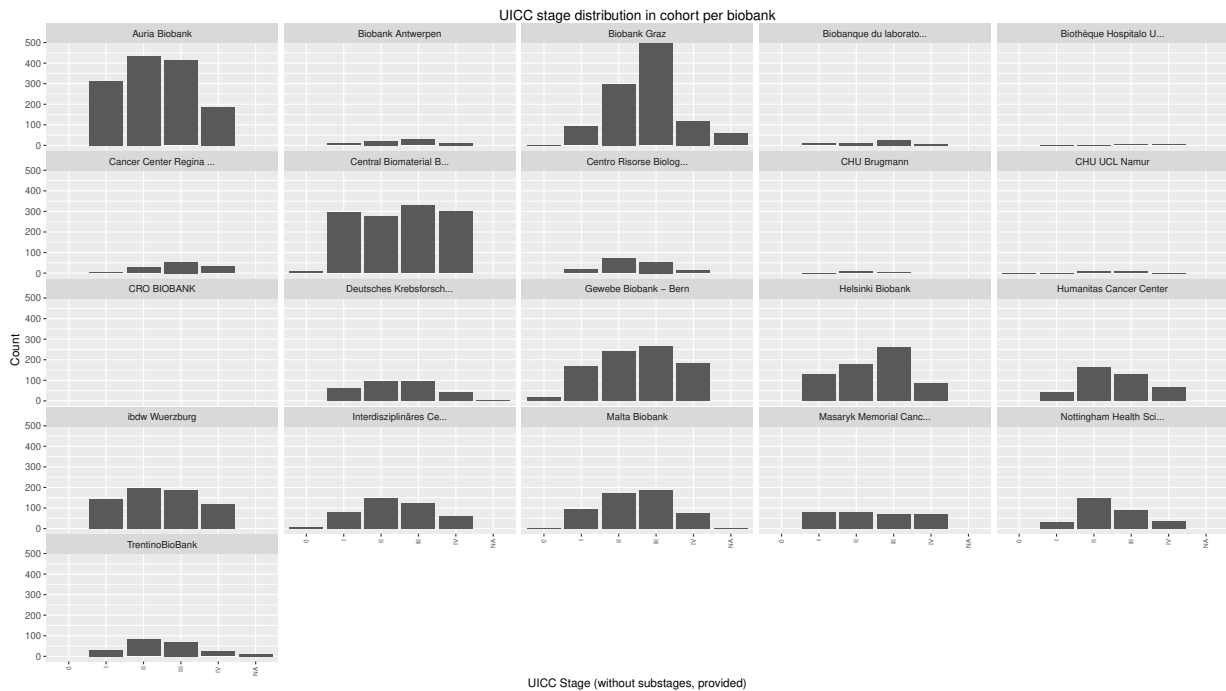


Figure 7: Per-biobank breakdown of distributions of available UICC stages as provided by biobanks.

Major differences

Biobank	Cases
Biobank Graz	32
Biobanque du laborato...	27
Centro Risorse Biolog...	2
CHU UCL Namur	2
Gewebe Biobank - Bern	3
ibdw Wuerzburg	5
Malta Biobank	22
TrentinoBioBank	15

Minor differences

Biobank	Cases
Auria Biobank	606
Biobank Antwerpen	3
Biobank Graz	48
Biobanque du laborato...	29
Central Biomaterial B...	46
Centro Risorse Biolog...	7
CHU UCL Namur	5
Gewebe Biobank - Bern	21
Humanitas Cancer Center	1
ibdw Wuerzburg	120
Interdisziplinäres Ce...	4
Malta Biobank	80
TrentinoBioBank	51

Table 3: Discrepancies between UICC stage as provided by biobanks, and values computed from pTNM values provided by biobanks according to the UICC standard. Major differences mean that the difference is at least on stage (e.g., II vs. III), while minor differences mean that there is a difference in sub-stage (e.g., IIA vs. IIB).



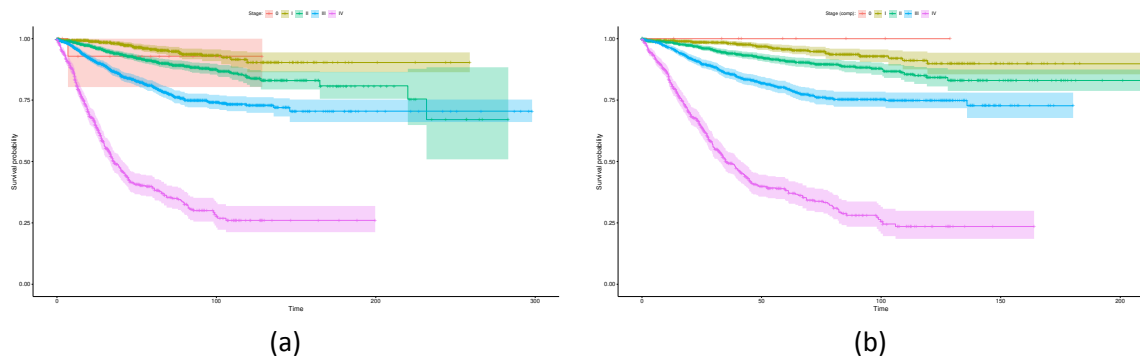
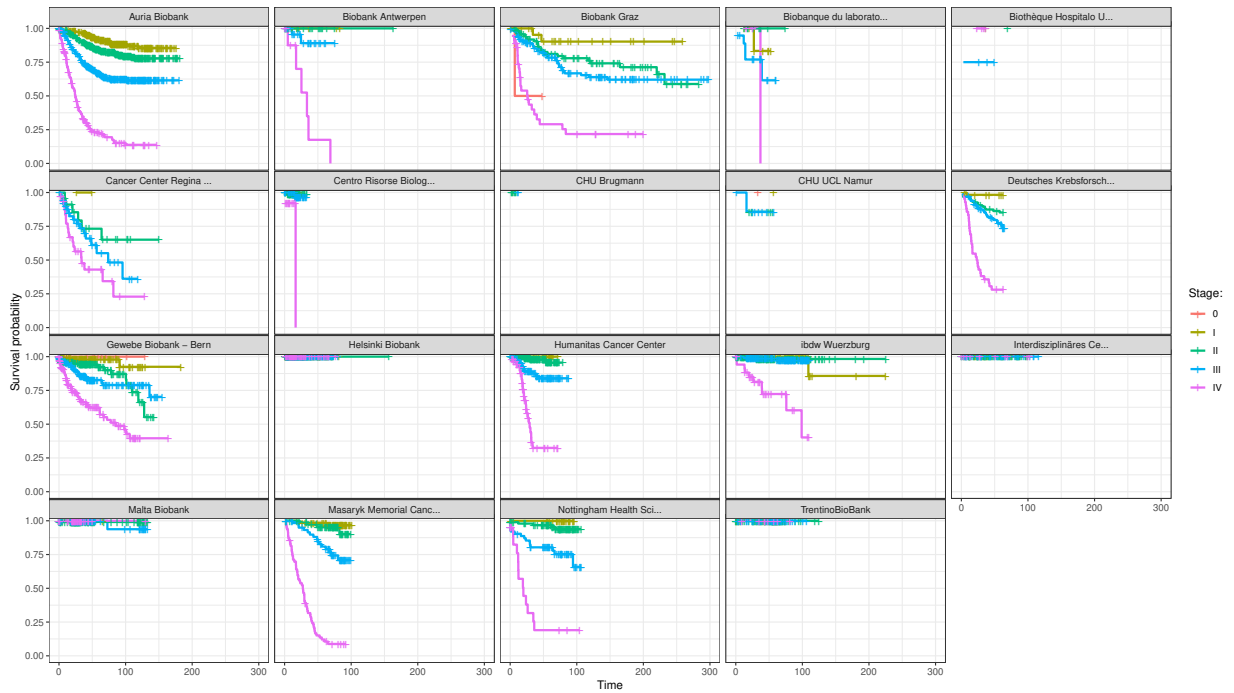
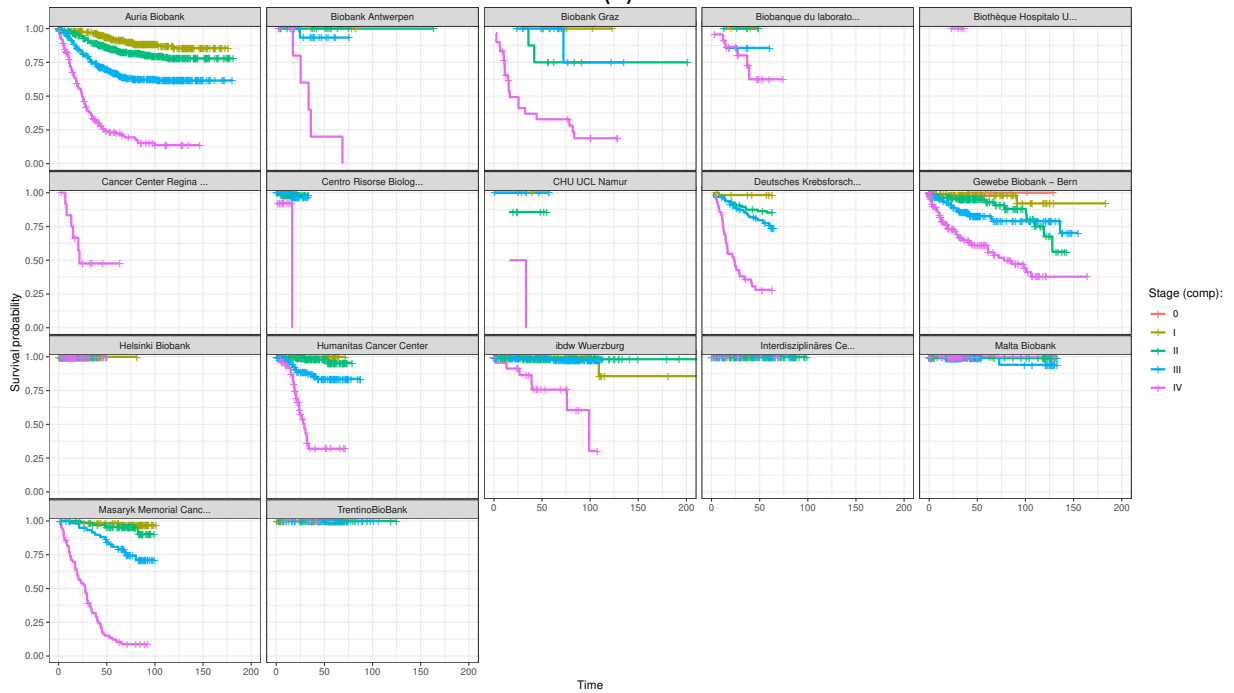


Figure 8: Comparison of survival curves for all cases between (a) provided UICC stages, and (b) stages computed from pTNM values. Time on x axis is months since initial diagnosis.



(a)



(b)

Figure 9: Decomposition of survival curves per biobank (a) based on provided UICC stages, (b) based on stages computed from pTNM values. Time on x axis is months since initial diagnosis.



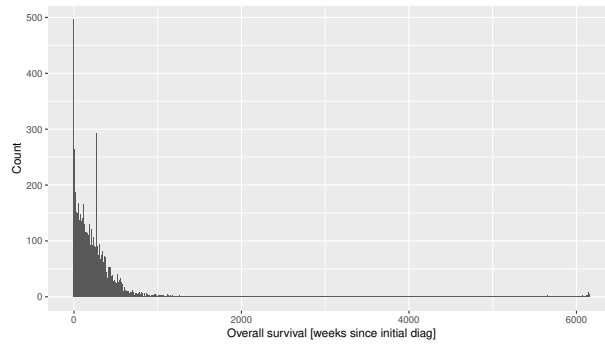


Figure 10: Overall survival in weeks since initial diagnosis.

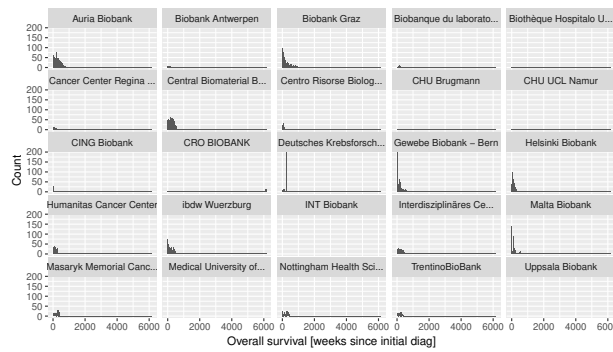


Figure 11: Per-biobank decomposition of overall survival in weeks since initial diagnosis.

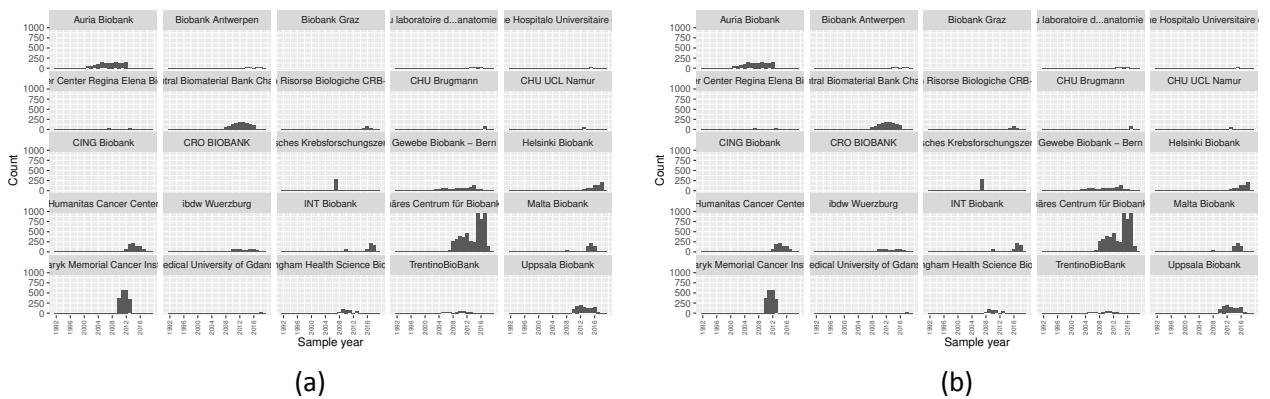


Figure 12: Per-biobank decomposition of (a) year of sample collection and (b) initial year of sample collection for each given patient.



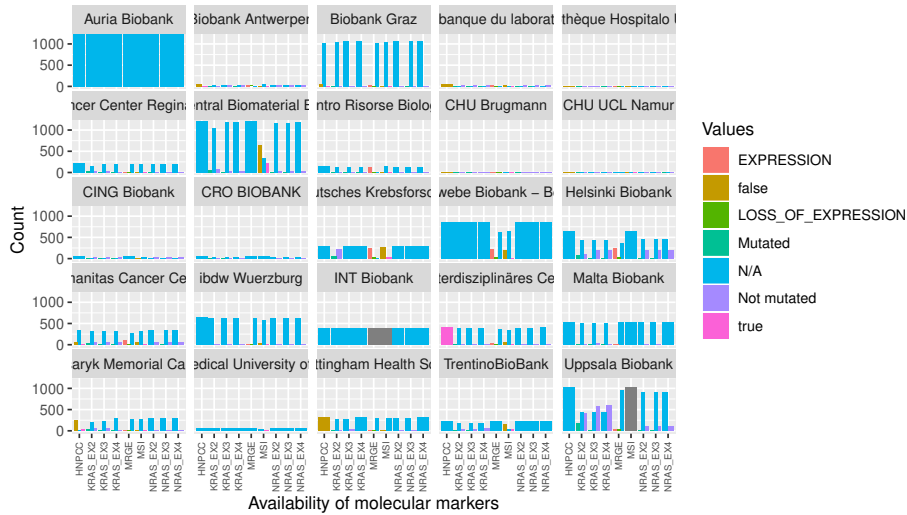


Figure 13: Availability of molecular markers per biobank.

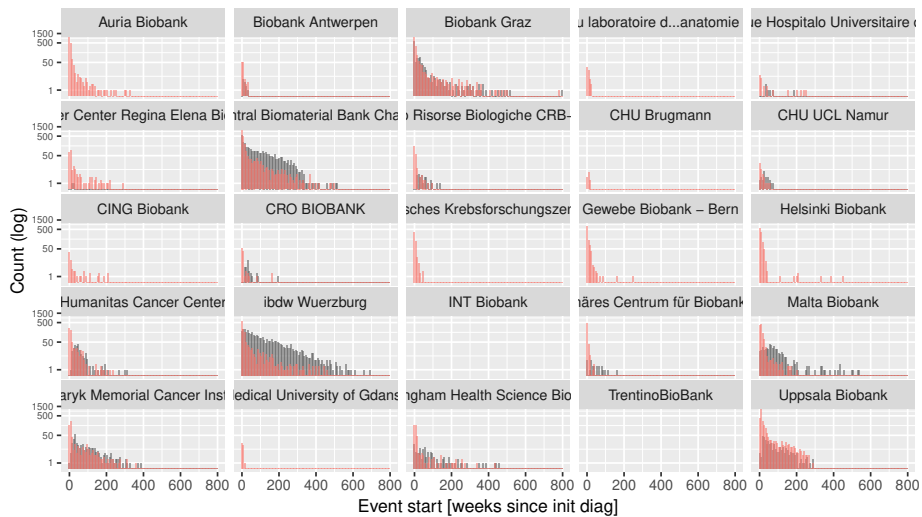


Figure 14: Per-biobank timeline of available events. The red columns represent therapy events and the gray columns responses to therapies.



3. Discussion and Conclusions

The ADOPT project has succeeded in reaching its goal to generate the CRC-Cohort of more than 10 000 cases, with truly pan-European coverage of the cohort; the actual number of collected cases is 10 480. 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 IT expertise to manipulate the data at source. Technical and organizational measures to ensure data security and protect privacy of the persons contributing their data to CRC-Cohort were in-depth discussed and the agreement on data transfer was reached in compliance with the General Data Protection Regulation (GDPR), with the IMI Code of Practice on Secondary Use of Medical Data in Scientific Research Projects and taking in consideration the differences in regulatory and ethical issues within the different European Countries.

3.1. Lessons learned.

The following lessons learned have been extracted by BBMRI-ERIC team coordinating implementation of the CRC-Cohort; many of them were anticipated and experimentally confirmed, but yet created issues because of the sheer complexity of the ADOPT consortium and because of number of contributing data sources.

- **Initiation of the data collection process requires at least 6–12 months time period** that needs to be accounted for as a part of the project planning, in particular if working with institutions with which the contractual situation needs to be newly set up.
- The institution responsible for the data collection needs to have an **experienced lawyer and experienced IT professional working together** on the data protection and contracting aspects. This group should prepare a good reasoning why a common contract structure should be used and push the contributing institutions to minimize requested deviations from that contract structure in order to have a homogeneous cohort.
- **When building a permanent resource** it is highly advisable for the **responsible institution to become also a data controller** (second in the row). An example of the contract is the CRC-Cohort Data Provider Agreement template, which is an appendix to the CRC-Cohort Data Protection Policy. The reason for this is that controller-processor relation often assumes time-limited duration of the contract and thus the asset ceases to exist after this period expires.
- Organization of such effort must assume **country-specific issues** that need to be handled specifically. In case of our CRC-Cohort, there were two particular issues identified:
 - Because of their national regulatory framework, **Finland does not allow to transfer data controllership outside of the country**. For this reason Finnish data is not part of many world-leading databases including dbSNP. In case of the CRC-Cohort, the problem was circumvented by developing a specific contract only for the duration of the data collection and quality checking process and then relying on making the data available via federated search only. This prac-



tically meant that the Finnish data was deleted from the CRC-Cohort after data quality checks were complete and the biobanks were asked to make the data available via a Connector for the federated search by the BBMRI-ERIC Locator platform.

- Italy needed specific contracting clauses to confirm that in case of use of the CRC-Cohort data for commercial research, they would be part of the decision process (which they would be anyway because of how the CRC-Cohort access procedure was set up, but still they needed a specific confirmation of this fact in the contract).
- **Approval of the contracts by the contributing institutions can takes between weeks and months**, even if the responsible institution has a quickly reacting lawyer. The reason is possibly several levels of approval needed, going through the data protection officer approval, going via ethics assessment etc. Hence also any subsequent changes to the contract, namely if induced by external reasons such as change of the legal framework on European or national level, shall be done very carefully to minimize risk of hampering the ongoing data collection process.
- At the time of implementing the ADOPT project, the **automated extraction of structured data from unstructured clinical records was not mature enough** to achieve reliable extraction of data across different European languages. This was further augmented by the fact that the project aimed at relatively broad geographical distribution of cases across Europe; hence many institutions were involved contributing low to medium numbers of cases per institution and thus significant customization or training of adaptive tools would not result in effective extraction process, since the training data sets needed would be at least as large as the extracted data sets.
- It is important to remind the source institutions that they can have **access to different national registries**, which contain volumes of structured data conveniently complementing the data available in the hospital or in the biobank. These range from national death registries to various disease-specific registries for major disease groups (typically some kind of national oncology registry).
- One often observes Pareto (80-20) distribution (of course as a very crude approximation) when it comes to data extraction: **extracting 80% of the data takes 20% of time** (typically structured data available in the biobank or hospital, which only needs to be converted to the target format, in our case it was), while **obtaining remaining 20% data takes 80% of time and funding resources** (in our case the most problematic parts were treatments and responses to treatments; this is because biobanks mostly used data from common clinical practice and not from the clinical trials, where the data would be already well-structured).
- The collection process also revealed strong need for **data quality checks**, which poses an interesting challenge for federated research infrastructures like BBMRI-ERIC, where the data stays primarily in the source repositories and are not available for central quality checks unless there are projects collecting the data centrally, like ADOPT did. **Biobanks should be also contractually bound to improve the data quality** if substantial defects are identified.
- Data **quality checking results raise substantial concerns for federated data querying systems**, such as BBMRI-ERIC Locator, where the source data must stay in the biobanks. The data quality checks can (shall) be also distributed as a part of the federated architecture, yet the central entity (in our case BBMRI-ERIC) has limited control of how these are handled and if the problems are indeed fixed or just not analyzed or not reported.



- It needs to be recognized that the **data quality improvement is a continuous process**. Even a simple static data quality checking may be an iterative process, as the updated data may result in additional quality-related findings that need to be corrected. Sometimes the data quality checks need to be further developed, too, e.g., based on detecting statistical properties of the whole data set that are not obvious with smaller portions of contributed data (e.g., comparisons of survival curves between different contributing institutions). **Biobanks need to be warned about the data quality improvement cycles** so that they allocate their resources accordingly and do not deplete all their resources on just the primary data collection.

All of these lessons learned will be taken into considerations when improving organization of BBMRI-ERIC biobanks and when designing and developing IT tools to make the biological material and data compliant with FAIR and FAIR-Health principles [2, 7].

The resulting CRC-Cohort is being made available for researchers in compliance with the access modes defined in the CRC-Cohort Data Protection Policy (see ADOPT Deliverable D2.3 Appendix III).



Bibliography

- [1] European Commission. “COMMISSION IMPLEMENTING DECISION of 22 November 2013 on setting up the Biobanks and Biomolecular Resources Research Infrastructure Consortium (BBMRI-ERIC) as a European Research Infrastructure Consortium (2013/701/EU)”. In: *Official Journal of the European Union* L 320/63. November (2013), pp. 63–80. URL: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:320:0063:0080:en:PDF>.
- [2] P. Holub, F. Kohlmayer, F. Prasser, M. T. Mayrhofer, I. Schlünder, G. M. Martin, S. Casati, L. Koumakis, A. Wutte, Ł. Kozera, et al. “Enhancing reuse of data and biological material in medical research: From FAIR to FAIR-health”. In: *Biopreservation and biobanking* 16.2 (2018), pp. 97–105.
- [3] D. Kadioglu, B. Breil, C. Knell, M. Lablans, S. Mate, D. Schlue, H. Serve, H. Storf, F. Ückert, T. Wagner, et al. “Samplify. MDR-A Metadata Repository and Its Application in Various Research Networks.” In: *Studies in health technology and informatics* 253 (2018), pp. 50–54.
- [4] S. Mate, M. Kampf, W. Rödle, S. Kraus, R. Proynova, K. Silander, L. Ebert, M. Lablans, C. Schüttler, C. Knell, et al. “Pan-European Data Harmonization for Biobanks in ADOPT BBMRI-ERIC”. In: *Applied clinical informatics* 10.04 (2019), pp. 679–692.
- [5] M. Muscholl, M. Lablans, T. O. Wagner, and F. Ückert. “OSSE—open source registry software solution”. In: *Orphanet journal of rare diseases* 9.S1 (2014), O9.
- [6] J. Schaaf, D. Kadioglu, J. Goebel, C. Behrendt, M. Roos, F. Ückert, F. Sadiku, T. Wagner, H. Storf, et al. “OSSE Goes FAIR-Implementation of the FAIR Data Principles for an Open-Source Registry for Rare Diseases.” In: *Studies in health technology and informatics* 253 (2018), pp. 209–213.
- [7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3 (2016).



A. XML Format for CRC-Cohort data import into Colorectal Cancer Data Collection system (CCDC)

This appendix provides the documentation that has been sent to the participating biobanks in January 2018, together with the examples of XML format in which they can deliver the data. The documentation provides a complete overview of the data model and its mapping to the XML entities used to import the data into the CCDC platform.



XML Format for CRC-Cohort data import into CCDC

BBMRI-ERIC ADOPT WP3 team

Monday 22nd January, 2018



Contents

1	Introduction	4
2	Patient Pseudonyms	5
3	Forms	6
3.1	form_28_ver-27 - Form	6
3.2	form_29_ver-5 - Form5	7
3.3	form_30_ver-3 - Form6	7
3.4	form_31_ver-2 - Form4	7
3.5	form_32_ver-8 - Form	8
3.6	form_33_ver-10 - Form3	8
3.7	form_34_ver-22 - Form2	8
3.8	form_35_ver-6 - Form1	9
4	Data Elements	10
4.1	Datelement_10_2 - PHARMACOTHERAPY_START_RELATIVE	11
4.2	Datelement_11_2 - PHARMACOTHERAPY_END_RELATIVE	12
4.3	Datelement_12_4 - RADIATION_THERAPY_START_RELATIVE	13
4.4	Datelement_13_2 - RADIATION_THERAPY_END_RELATIVE	14
4.5	Datelement_14_3 - MM_MICROSAT_INSTABILITY	15
4.6	Datelement_15_2 - MM_MISMATCH_REPAIR_GE	16
4.7	Datelement_16_3 - MM_RISK_SITUATION_HNPCC	17
4.8	Datelement_20_3 - MM_KRAS_MUTATION_KRAS_EX2	18
4.9	Datelement_21_5 - MM_KRAS_MUTATION_KRAS_EX3	19
4.10	Datelement_22_4 - MM_KRAS_MUTATION_KRAS_EX4	20
4.11	Datelement_23_5 - MM_KRAS_MUTATION_NRAS_EX2	21
4.12	Datelement_24_4 - MM_KRAS_MUTATION_NRAS_EX3	22
4.13	Datelement_25_3 - MM_KRAS_MUTATION_NRAS_EX4	23
4.14	Datelement_2_2 - CLINICAL_STUDY_PARTICIPANT	24
4.15	Datelement_30_3 - DIAG_MRI_DONE	25
4.16	Datelement_31_3 - DIAG_CT_DONE	26
4.17	Datelement_33_1 - THERAPY_RESPONSE	27
4.18	Datelement_34_1 - THERAPY_RESPONSE_TIMESTAMP_RELATIVE	28
4.19	Datelement_35_3 - TARGETED_THERAPY_START_RELATIVE	29
4.20	Datelement_36_1 - TARGETED_THERAPY_END_RELATIVE	30
4.21	Datelement_3_1 - AGE_AT_PRIMARY_DIAGNOSIS	31
4.22	Datelement_49_1 - SURGERY_TYPE	32
4.23	Datelement_4_3 - TIME_OF_RECURRENCE_RELATIVE	33
4.24	Datelement_51_3 - DATE_DIAGNOSIS	34
4.25	Datelement_53_3 - WHO_GRADE_VERSION	35
4.26	Datelement_54_2 - SAMPLE_MATERIAL_TYPE	36
4.27	Datelement_55_2 - SAMPLE_PRESERVATION_MODE	37



4.28	Dataelement_56_2 - SAMPLE_ID	38
4.29	Dataelement_57_3 - DIGITAL_IMAGING_AVAILABILITY	39
4.30	Dataelement_58_2 - DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY	40
4.31	Dataelement_59_5 - PHARMACOTHERAPY_SCHEME	41
4.32	Dataelement_5_2 - VITAL_STATUS	43
4.33	Dataelement_61_5 - DIAG_LIVER_IMAGING_DONE	44
4.34	Dataelement_63_4 - DIAG_X_DONE	45
4.35	Dataelement_67_1 - SURGERY_TYPE_OTHER	46
4.36	Dataelement_68_2 - HIST_METASTASIS	47
4.37	Dataelement_6_3 - VITAL_STATUS_TIMESTAMP	48
4.38	Dataelement_70_2 - UICC_STAGE	49
4.39	Dataelement_71_1 - TNM_PRIMARY_TUMOR	50
4.40	Dataelement_73_3 - UICC_VERSION	51
4.41	Dataelement_75_1 - TNM_DISTANT_METASTASIS	52
4.42	Dataelement_77_1 - TNM_REGIONAL_LYMPH_NODES	53
4.43	Dataelement_7_2 - OVERALL_SURVIVAL_STATUS	54
4.44	Dataelement_81_3 - PHARMACOTHERAPY_SCHEME_DESCRIPTION	55
4.45	Dataelement_82_1 - BIOLOGICAL_MATERIAL_FROM_RECURRENCE_AVAILABLE	56
4.46	Dataelement_83_1 - WHO_GRADE	57
4.47	Dataelement_85_1 - SEX	58
4.48	Dataelement_87_1 - BRAF_PIC3CA_HER_MUTATION_STATUS	59
4.49	Dataelement_88_1 - DIAG_COLONOSCOPY	60
4.50	Dataelement_89_3 - YEAR_OF_SAMPLE_COLLECTION	61
4.51	Dataelement_8_3 - SURGERY_START_RELATIVE	62
4.52	Dataelement_91_1 - HIST_MORPHOLOGY	63
4.53	Dataelement_92_1 - HIST_LOCALIZATION	65
4.54	Dataelement_93_1 - SURGERY_LOCATION	66
4.55	Dataelement_9_2 - SURGERY_RADICALITY	67



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 constraints 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,
2. 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.



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.



3 Forms

3.1 form_28_ver-27 - Form

Required data elements

- Dataelement_14_3 - MM_MICROSAT_INSTABILITY
- Dataelement_15_2 - MM_MISMATCH_REPAIR_GE
- Dataelement_20_3 - MM_KRAS_MUTATION_KRAS_EX2
- Dataelement_21_5 - MM_KRAS_MUTATION_KRAS_EX3
- Dataelement_22_4 - MM_KRAS_MUTATION_KRAS_EX4
- Dataelement_23_5 - MM_KRAS_MUTATION_NRAS_EX2
- Dataelement_24_4 - MM_KRAS_MUTATION_NRAS_EX3
- Dataelement_25_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_2 - VITAL_STATUS
- Dataelement_61_5 - DIAG_LIVER_IMAGING_DONE
- Dataelement_63_4 - DIAG_X_DONE
- Dataelement_7_2 - OVERALL_SURVIVAL_STATUS
- Dataelement_85_1 - SEX
- Dataelement_88_1 - DIAG_COLONOSCOPY

Optional data elements

- Dataelement_16_3 - MM_RISK_SITUATION_HNPCC
- Dataelement_2_2 - 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_12_4 - RADIATION_THERAPY_START_RELATIVE
- Dataelement_13_2 - 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_2 - 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_2 - PHARMACOTHERAPY_START_RELATIVE
- Dataelement_11_2 - 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_2 - HIST_METASTASIS
- Dataelement_70_2 - UICC_STAGE
- Dataelement_71_1 - TNM_PRIMARY_TUMOR
- Dataelement_73_3 - UICC_VERSION



- Dataelement_75_1 - TNM_DISTANT_METASTASIS
- Dataelement_77_1 - TNM_REGIONAL_LYMPH_NODES
- Dataelement_83_1 - WHO_GRADE
- Dataelement_91_1 - HIST_MORPHOLOGY
- Dataelement_92_1 - HIST_LOCALIZATION

Optional data elements

- Dataelement_57_3 - DIGITAL_IMAGING_AVAILABILITY
- Dataelement_58_2 - DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY
- Dataelement_82_1 - BIOLOGICAL_MATERIAL_FROM_RECURRENCE_AVAILABLE

Other data elements None.

3.8 form_35_ver-6 - Form1

Required data elements

- Dataelement_54_2 - SAMPLE_MATERIAL_TYPE
- Dataelement_55_2 - SAMPLE_PRESERVATION_MODE
- Dataelement_56_2 - 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.



4.1 Dataelement_10_2 - PHARMACOTHERAPY_START_RELATIVE

XSD label Dataelement_10_2

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] ($0 \leq x$)

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.



4.2 Dataelement_11_2 - PHARMACOTHERAPY_END_RELATIVE

XSD label Dataelement_11_2

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] ($0 \leq x$)

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.



4.3 Dataelement_12_4 - RADIATION_THERAPY_START_RELATIVE

XSD label Dataelement_12_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] ($0 \leq x$)

XSD parent form Form5 - form_29_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.



4.4 Dataelement_13_2 - RADIATION_THERAPY_END_RELATIVE

XSD label Dataelement_13_2

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] ($0 \leq x$)

XSD parent form Form5 - form_29_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.



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_28_ver-27

XSD description

Microsatellites analysed BAT26, D17S250, D5S346, BAT40, D2S123 and BAT25. Image cytometry does not qualify for comparability reasons

Data model description

Microsatellites analysed BAT26, D17S250, D5S346, BAT40, D2S123 and BAT25. Image cytometry does not qualify for comparability reasons



4.6 Dataelement_15_2 - MM_MISMATCH_REPAIR_GE

XSD label Dataelement_15_2

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_28_ver-27

XSD description

Mismatch repair gene expression – IHC array for different genes (common for 3). Expression of MLH1, MSH2, PMS2 and MSH6

Data model description

Mismatch repair gene expression – IHC array for different genes (common for 3). Expression of MLH1, MSH2, PMS2 and MSH6



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_28_ver-27

XSD description

Risk situation (only HNPCC), Amsterdam criteria

Data model description

Risk situation (only HNPCC), Amsterdam criteria



4.8 Dataelement_20_3 - MM_KRAS_MUTATION_KRAS_EX2

XSD label Dataelement_20_3

Data model label MM_KRAS_MUTATION_KRAS_EX2

Level in data model REQUIRED

XSD name KRAS exon 2 (codons 12 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_28_ver-27

XSD description

KRAS exon 2 (codons 12 or 13) mutation status

Data model description

KRAS exon 2 (codons 12 or 13) mutation status



4.9 Dataelement_21_5 - MM_KRAS_MUTATION_KRAS_EX3

XSD label Dataelement_21_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_28_ver-27

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_22_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_28_ver-27

XSD description

KRAS exon 4 (codons 117 or 146)

Data model description

KRAS exon 4 (codons 117 or 146)



4.11 Dataelement_23_5 - MM_KRAS_MUTATION_NRAS_EX2

XSD label Dataelement_23_5

Data model label MM_KRAS_MUTATION_NRAS_EX2

Level in data model REQUIRED

XSD name NRAS exon 2 (codons 12 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_28_ver-27

XSD description

NRAS exon 2 (codons 12 or 13) mutation status

Data model description

NRAS exon 2 (codons 12 or 13) mutation status



4.12 Dataelement_24_4 - MM_KRAS_MUTATION_NRAS_EX3

XSD label Dataelement_24_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_28_ver-27

XSD description

NRAS exon 3 (codons 59 or 61) mutation status

Data model description

NRAS exon 3 (codons 59 or 61) mutation status



4.13 Dataelement_25_3 - MM_KRAS_MUTATION_NRAS_EX4

XSD label Dataelement_25_3

Data model label MM_KRAS_MUTATION_NRAS_EX4

Level in data model REQUIRED

XSD name NRAS exon 4 (codons 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_28_ver-27

XSD description

NRAS exon 4 (codons 117 or 146) mutation status

Data model description

NRAS exon 4 (codons 117 or 146) mutation status



4.14 Dataelement_2_2 - CLINICAL_STUDY_PARTICIPANT

XSD label Dataelement_2_2

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_28_ver-27

XSD description

Participation in clinical study

Data model description

Participation in clinical study



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_28_ver-27

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.



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_28_ver-27

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.



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-2

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 [] ($0 \leq x$)

XSD parent form Form4 - form_31_ver-2

XSD description

Date response was obtained in weeks since initial diagnosis

Data model description

Date response was obtained in weeks since initial diagnosis



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 [] ($0 \leq x$)

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)



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 [] ($0 \leq x$)

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)



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] ($0 \leq x$)

XSD parent form Form - form_28_ver-27

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.



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_32_ver-8

XSD description

Surgery type

Data model description

Surgery type



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] ($0 \leq x$)

XSD parent form Form - form_28_ver-27

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



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_28_ver-27

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



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 Form2 - form_34_ver-22

XSD description

The version of the WHO classification system used. Version years: 4th ed. (used since 2011), 3rd ed. (2001-2010), 2nd ed. (1991-2000), 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_2

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



4.27 Dataelement_55_2 - SAMPLE_PRESERVATION_MODE

XSD label Dataelement_55_2

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



4.28 Dataelement_56_2 - SAMPLE_ID

XSD label Dataelement_56_2

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



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 Form2 - form_34_ver-22

XSD description

Do you have high-resolution digital imaging (corresponding to magnification 40x) from the histopathology?. 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 histopathology?. 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).



4.30 Dataelement_58_2 - DIGITAL_IMAGING_INVASION_FRONT_AVAILABILITY

XSD label Dataelement_58_2

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 Form2 - form_34_ver-22

XSD description

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

Data model description

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



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/



- ↪ 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 therapy 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 therapy 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_2

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_28_ver-27

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_28_ver-27

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_28_ver-27

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.



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_32_ver-8

XSD description

Surgery type, if not present on the list

Data model description

Surgery type, if not present on the list



4.36 Dataelement_68_2 - HIST_METASTASIS

XSD label Dataelement_68_2

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 Form2 - form_34_ver-22

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_28_ver-27

XSD description

Timestamp of last update of vital status

Data model description

Timestamp of last update of vital status



4.38 Dataelement_70_2 - UICC_STAGE

XSD label Dataelement_70_2

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 Form2 - form_34_ver-22

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 Form2 - form_34_ver-22

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



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 Form2 - form_34_ver-22

XSD description

The version of the UICC system under which the staging was done. Version years:8th edition (since 2017),7th edition (used 2010-2017),6th. ed (used 2003-2009),5th. ed (used 1998-2002),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 Form2 - form_34_ver-22

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



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 Form2 - form_34_ver-22

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



4.43 Dataelement_7_2 - OVERALL_SURVIVAL_STATUS

XSD label Dataelement_7_2

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] ($0 \leq x$)

XSD parent form Form - form_28_ver-27

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



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



4.45 Dataelement_82_1 - BIOLOGICAL_MATERIAL_FROM_RECURRENCE_AVAILABLE

XSD label Dataelement_82_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 Form2 - form_34_ver-22

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 Form2 - form_34_ver-22

XSD description

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

Data model description

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



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_28_ver-27

XSD description

Biological sex of the person, defined by chromosomes.

Data model description

Biological sex of the person, defined by chromosomes.



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, HER2 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_28_ver-27

XSD description

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

Data model description

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



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_28_ver-27

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.



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] ($0 \leq x$)

XSD parent form Form1 - form_35_ver-6

XSD description

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

Data model description

Calendar year in which the sample was collected.



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] ($0 \leq x$)

XSD parent form Form - form_32_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.



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 Form2 - form_34_ver-22



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



4.53 Dataelement_92_1 - HIST_LOCALIZATION

XSD label Dataelement_92_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 Form2 - form_34_ver-22

XSD description

Histopathology part - Localization of primary tumor

Data model description

Histopathology part - Localization of primary tumor



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_32_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_2

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_32_ver-8

XSD description

Whether the surgery removed the entire tumor.

Data model description

Whether the surgery removed the entire tumor.

