ADOPT BBMRI-ERIC
gateway for health

ADOPT BBMRI-ERIC
GRANT AGREEMENT NO. 676550

DELIVERABLE REPORT

Deliverable no	2.5
Deliverable Title	Data matrix for characterisation of diseases
Contractual delivery month	M24
Responsible Partner	BBMRI.it/UNIMIB
Author(s)	Marialuisa Lavitrano, Georges Dagher

TITLE

Data matrix for characterisation of diseases

Executive Summary

We developed a general model for identification, selection and collection of samples and data from National Nodes that can be adapted to a broad spectrum of diseases and we defined a general matrix for characterization of common complex diseases. We also defined the properties that are necessary for unambiguous interpretation of the results. The goal was to provide a detailed description of the data structure in order to (a) create useful data set for each disease, (b) allow for unambiguous interpretation of the data when used in the research.

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

ADOPT BBMRI-ERIC gateway for health

Copyright notice

This work by Parties of the ADOPT BBMRI-ERIC Consortium is license under a Creative Commons Attributation 4.0 Internaltional License (http://creativecommons.org/licenses/by/4.0/).

Background

With the arrival of the post-genomic era, the use of the candidate gene approach – in which gene candidacy is estimated based on knowledge about gene function – to augment or even supplant positional cloning has become more prevalent, and this trend will probably continue into the future. Disease phenotypes provide a valuable window into the function of genes, which can be exploited in the candidate gene approach. The aim of ADOPT is to provide the matrix to build a minimum data base which will be used to identify and locate samples from patients with common complex diseases.
A consensus has been reached on the final data set to be collected for colon cancer (D2.4). This data set comprises required criteria for a correct interpretation of the results. The availability of the required data was assessed in the 75 participants biobanks. Datasets from more than 10.000 patients were available. Based on these encouraging results and the availability of these data in the biobanks, we developed a general model for identification, selection and collection of samples and data from National Nodes that can be adapted to a broad spectrum of diseases and we defined a general matrix for characterization of common complex diseases.

Approaches (Methods)

The development of a general matrix for characterization of common complex diseases was done in five steps. 1. Establishment of a multidisciplinary medical working group from BBMRI.it and BBMRI.fr including clinicians, surgeons, pathologists, radiologists, geneticists and bio-informaticians. 2. Definition of criteria to transfer the methodology set up for colorectal cancer dataset/sample collection to other disease entities. 3. Achievement of consensus on items to be included in the matrix among medical experts of BBMRI.it and of BBMRI.fr. 4. Review of the matrix by the joint group of multidisciplinary medical working group from BBMRI.it and BBMRI.fr. 5. Approval of the resulting matrix by the BBMRI-ERIC Management Committee #18 (October 2017).

Results

To transfer the methodology set up for colorectal cancer dataset/sample collection to other disease entities we define the following steps:

- Identify biobanks willing to contribute to the collection of samples/dataset for a specific disease by sending a questionnaire/survey to all National Node directors to map qualified BBs
- Establish the disease-specific dataset: a group of experts define a minimum list of data items
- Selection of samples/data: samples/data shall comply to quality, ethical and legal requirements

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

- Invitation to contribute with samples/data to all qualified BBs
- Collection of samples/data in a standardised format

We defined a general matrix for common complex diseases. This includes at least the following items:

- Disease classification (ICD code)
- Personal data (data of birth, sex, ethnicity, age at diagnosis, familial incidence)
- Physical examination (BP, weight, BM…)
- Life style data/risk factors (smoking, alcohol, fat intake… / direct causality : polyps, hypertension, obesity…)
- Familial history
- Comorbidity: Charlson index
- Performance scale: Karnofsky scale, ECOG
- Laboratory data
- Molecular signature
- Histopathology
- Imaging/instrumental data (X-ray, PET, Scan, MRI, colonoscopy, endoscopy…)
- Treatment/ Treatment response / other treatments
- Surgery
- Follow up data/recurrence
- Survival

These items have to be specified for each of the common diseases to be included in the database. It constitutes a minimum data set that will enable to collect samples from patients with similar inclusion criteria that will be used for genomic and other -omics studies.
A consensus has been reached on the definition of the Data Set items on the basis of the Data Model for Colorectal Cancer Gathering (**Appendix I**). All the properties that are necessary for unambiguous interpretation of the results have been defined. This includes the following properties for each collected variable (attribute):

- Unique label of the variable.
- Short description (label) of the variable - to be used in forms.
- Semantics = definition of meaning.
  - This includes references to existing clearly defined official standards or community "standards", including existing ontologies.
- Syntax
  - including data type (elementary types such as boolean, float, integer, free text, specifically structured text, etc., array or lists of elementary types),
  - including coding (e.g., IEEE 754 for floats, regular expressions for structured text).
- List of allowed units
  - including their conversion algorithms (with "non-existent" and "unknown" interim options)
- Level: REQUIRED, OPTIONAL, RECOMMENDED
  - REQUIRED means the data can't be entered at all without this item being provided,
  - OPTIONAL means data may or may not be provided, but the item will be ready for inputting the data in as part of the data model,

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

  - RECOMMENDED is a special subclass of the OPTIONAL, which is highly-recommended to be filled in (intended for items where we need the data but where we know that some sources won't be able to fill this in and we still want such data not being discarded as invalid).
- Relation to entities (patient, examination, etc.) - will be used for developing the formal model.

Each single item of the dataset is identified based upon the criteria defined. The goal is to provide a detailed description of the data structure in order to (a) create useful data set for each disease, (b) allow for unambiguous interpretation of the data when used in the research.

Discussion and Conclusions

Different bio-informatic strategies have been developed to predict and prioritize potential disease genes, and several web tools and bioinformatic approaches are now available. They can be classified into three broad categories: those based on intrinsic disease gene properties, those that use expression patterns and phenotypic information directly, and those that look at functional relatedness between candidate genes. These three categories differ in the extent to which they utilize disease phenotype information varying from simply distinguishing between disease and non-disease phenotypes on the one hand to the systematic comparison of disease phenotypes at the individual trait level on the other. The present matrix when implemented for specific diseases would foster access to samples and their analysis thus contributing to the deciphering candidate genes and pathways underlying common complex diseases.

Next Steps

Implementation of the matrix.

Appendices

Appendix I: Data Model for CRC Data Gathering

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

ADOPT BBMRI-ERIC gateway for health

Appendix I: Data Model for CRC Data Gathering

Definition of Data Model for Colorectal Cancer Data Gathering in ADOPT BBMRI-ERIC

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

PURPOSE OF THIS DOCUMENT

This document is intended to define the data set to be collected as a part of ADOPT BBMRI-ERIC project. The goal is to provide a detailed description of the data structure in order to (a) create a useful data set for the colon cancer research, (b) allow for *unambiguous interpretation* of the data when used in the research. The development will be done in several steps:

1. Basic consensus on collected attributes among the medical experts.
   Deadline: April 30, 2016
2. Development of formal model including entities, their attributes and their mutual relations by IT experts.
3. Review of the formal model by the joint group of medical and IT experts.
4. Approval of the resulting formal model by the BBMRI-ERIC Management Committee (used also as project management board in ADOPT BBMRI-ERIC project).
   Deadline: June 30, 2016
5. Development of the data collection application (ADOPT WP3) for manual data collection (manual data collection itself will be done within ADOPT WP2).
   Deadline: September 30, 2016.

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

ADOPT BBMRI-ERIC gateway for health

ORIGINAL DATA SET PROPOSED IN ADOPT BBMRI-ERIC

The following general data set is proposed in ADOPT BBMRI-ERIC WP2 Task 3 - we are now seeking to make it more precise.

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

Inclusion criteria

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

- Colorectal cancer as a primary diagnosis (C18.1 to C18.7, C19, C20)
- Available FFPE – surgical material
- Availability of all REQUIRED data
- Willingness to provide access to (a) samples, (b) pseudonymized data as a part of (i) participation in research projects, (ii) cost or no-cost recovery procedure. This assumes

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

signing MTA/DTA.

DEFINITION OF THE ITEMS

How items need to be defined

We need to define all the properties that are necessary for unambiguous interpretation of the results. This includes the following properties for each collected variable (attribute):

- Unique label of the variable.
- Short description (label) of the variable - to be used in forms.
- Semantics = definition of meaning.
  - This includes references to existing clearly defined official standards or community "standards", including existing ontologies.
  - We will use this for ontologizing the data model, in order to make it "machine readable" (allowing for correct interpretation of the data in automated processing workflows).
- Syntax
  - including data type (elementary types such as boolean, float, integer, free text, specifically structured text, etc., array or lists of elementary types),
  - including coding (e.g., IEEE 754 for floats, regular expressions for structured text).
- List of allowed units
  - including their conversion algorithms (with "non-existent" and "unknown" interim options)
- Level: REQUIRED, OPTIONAL, RECOMMENDED
  - REQUIRED means the data can't be entered at all without this item being provided,
  - OPTIONAL means data may or may not be provided, but the item will be ready for inputting the data in as part of the data model,
  - RECOMMENDED is a special subclass of the OPTIONAL, which is highly-recommended to be filled in (intended for items where we need the data but where we know that some sources won't be able to fill this in and we still want such data not being discarded as invalid).
- Relation to entities (patient, examination, etc.) - will be used for developing the formal model.

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

ADOPT BBMRI-ERIC gateway for health

Defined variables:

- Sex:
  - Label: SEX
  - Short description: Biological sex
  - Semantics:
    - Biological sex of the person, defined by chromosomes.
    - http://purl.obolibrary.org/obo/PATO_0020000
  - Syntax: male, female (only 2 values allowed)
  - Units: n/a
  - Level: REQUIRED
- Participation in clinical study
  - 1..1 to patient
  - Label: CLINICAL_STUDY_PARTICIPANT
  - Semantics:
    - Participant of clinical study.
    - TODO: link to ontology
  - Syntax: boolean
  - Units: n/a
  - Level: RECOMMENDED
- Age at primary diagnosis:
  - Label: AGE_AT_PRIMARY_DIAGNOSIS
  - Short description: Age at diagnosis (rounded to years)
  - Semantics:
    - Age at initial histopathological diagnosis (biopsy or surgical specimen of the primary tumor) rounded to years.
    - http://purl.bioontology.org/ontology/SNOMEDCT/423493009
  - Syntax: integer
  - Units: years since birth
  - Level: REQUIRED
- Time of recurrence (metastasis):
  - 0..n related to the patient
  - Label: TIME_OF_RECURRENCE_RELATIVE
  - Short description: Time of recurrence (metastasis diagnosis)
  - Semantics:
    - Weeks between primary diagnosis (AGE_AT_PRIMARY_DIAGNOSIS) and diagnosed recurrence
    - If only months is available, conversion is weeks := months * 4

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

        - TODO - link to ontology (Esmeralda - to TCGA)
    - Syntax: integer
    - Units: weeks since primary diagnosis
    - Level: OPTIONAL
- Vital status and survival information o 1..1 to person (= REQUIRED) o Vital status
        - Label: VITAL_STATUS
        - Semantics: living or deceased
        - Syntax:
            - list (ALIVE= … person is still alive, DEATH_COLON_CANCER = death due to colon cancer, DEATH_OTHER = death due to other reasons, DEATH_UNKNOWN_REASON = death for unknown reasons, UNKNOWN = unknown)
            - TODO: link to ontology - Esmeralda
        - Units: n/a
        - Level: REQUIRED
    - Timestamp of last update of vital status
        - Label: VITAL_STATUS_TIMESTAMP
        - Semantics:
            - Timestamp of last update of vital status
            - TODO: link to ontology - Esmeralda
        - Syntax: timestamp compliant to ISO 8601
        - Units: n/a
        - Level: REQUIRED if VITAL_STATUS != UNKNOWN
    - Overall survival status
        - Label: OVERALL_SURVIVAL_STATUS
        - Semantics:
            - Weeks after first colon cancer therapy started for the given person.
            - If the data is collected at the source in months only, the conversion should be weeks := months*4
            - TODO: link to ontology - Esmeralda
        - Syntax: integer
        - Units: weeks
        - Level: REQUIRED
- Surgery: aggregate object

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

- 0..n - patient to surgery
- Time difference between initial diagnosis and surgery:
  - Label: SURGERY_START_RELATIVE
  - Semantics:
    - Weeks between initial diagnosis and date of surgery.
    - TODO: link to ontology
  - Syntax: integer
  - Units: weeks
  - REQUIRED
- Surgery radicality:
  - Label: SURGERY_RADICALITY
  - Semantics:
    - Whether the surgery removed the entire tumor.
    - TODO: link to ontology
  - Syntax: list (RX, R0, R1, R2)
  - Units: n/a
  - Level: REQUIRED
- Type of surgery:
  - Label: SURGERY_TYPE
  - Semantics:
    - "OTHER" value may allow for optional "please specify" free text option
    - TODO: link to ontology
  - Syntax: list (RIGHT_HEMICOLECTOMY, LEFT_HEMICOLECTOMY, TRANSVERSE_COLECTOMY, SIGMOID_COLECTOMY, TOTAL_COLECTOMY, PAN- PROCTO_COLECTOMY, LOW_ANTERIOR_COLON_RESECTION, ANTERIOR_RESECTION_OF_RECTUM, ABDOMINO-PERINEAL_RESECTION, ENDO- RECTAL_TUMOR_RESECTION, OTHER)
  - Units: n/a
  - Level: REQUIRED

- Pharmacotherapy:
  - 0..n to patient
  - REQUIRED if occurred
  - Date of start:
    - Label: PHARMACOTHERAPY_START_RELATIVE
    - Semantics:
      - start of the drug intake in weeks since initial diagnosis.
      - TODO: link to ontology
    - Syntax: integer
    - Units: weeks

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

        - Level: REQUIRED
    - Date of end:
        - Label: PHARMACOTHERAPY_END_RELATIVE
        - Semantics:
            - end of the drug intake in weeks since initial diagnosis.
            - TODO: link to ontology
        - Syntax: integer
        - Units: weeks
        - Level: REQUIRED
    - Scheme of pharmacotherapy:
        - Label: PHARMACOTHERAPY_SCHEME
        - Semantics:
            - Pointer to one of the rows of the following table:

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

UFT, uracil–tegafur.

        - Syntax: list (rows from the table above)
        - Units: N/A
        - Level: REQUIRED
- Targeted therapy:
    - 0..n to patient
    - REQUIRED if occurred
    - Date of start:
        - Label: TARGETED_THERAPY_START_RELATIVE
        - Semantics:
            - start of the drug intake in weeks since initial diagnosis.
            - TODO: link to ontology
        - Syntax: integer
        - Units: weeks

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

        - Level: REQUIRED
    - Date of end:
        - Label: TARGETED_THERAPY_END_RELATIVE
        - Semantics:
            - end of the drug intake in weeks since initial diagnosis.
            - TODO: link to ontology
        - Syntax: integer
        - Units: n/a
        - Level: OPTIONAL
- Radiation therapy:
    - 0..1 to patient
        - REQUIRED if occurred
        - Date of start:
            - Label: RADIATION_THERAPY_START_RELATIVE
            - Semantics:
                - start of the radiation therapy in weeks since initial diagnosis.
                - TODO: link to ontology
            - Syntax: integer
            - Units: weeks
            - Level: REQUIRED
        - Date of end:
            - Label: RADIATION_THERAPY_END_RELATIVE
            - Semantics:
                - end of the radiation therapy in weeks since initial diagnosis.
                - TODO: link to ontology
            - Syntax: integer
            - Units: weeks
            - Level: REQUIRED
    - Response to therapy
        - 0..n to patient
        - The response is linked to the patient and specified by a timestamp. This is to avoid need to specify to which therapy the response is, since there might be combination of different therapies.
        - Specific response
            - Label: THERAPY_RESPONSE
            - Semantics:
                - Therapy response according to RECIST criteria[1].
                - TODO: link to ontology
            - Syntax: LIST (PROGRESSIVE_DISEASE, STABLE_DISEASE, PARTIAL_RESPONSE, COMPLETE_RESPONSE)

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

        - Units: n/a
        - Level: REQUIRED (only if the response exists - see overall 0..n relation to patient)
    - Specific response
        - Label: THERAPY_RESPONSE_TIMESTAMP_RELATIVE
        - Semantics:
            - Timestamp when the therapy response was obtained, in weeks relative to the initial diagnosis
            - Weeks := months * 4 (if only months are available)
            - TODO: link to ontology
        - Syntax: integer
        - Units: weeks
        - Level: REQUIRED (only if the response exists - see overall 0..n relation to patient)
- Molecular markers
    - Microsatellite instability
        - 1..1 to person
        - Label: MM_MICROSAT_INSTABILITY
        - Semantics:
            - Microsatellites analysed BAT26, D17S250, D5S346, BAT40, D2S123 and BAT25
        - Syntax: LIST (NOT_DONE, NO, YES) – SINGLE-VALUE
        - Units: N/A
        - Level: REQUIRED
    - Mismatch repair gene expression – IHC array for different genes (common for 3)
        - 1..1 to person
        - Label: MM_MISMATCH_REPAIR_GE
        - Semantics:
            - existing guidelines - immunohistochemistry
            - Expression of MLH1, MSH2, PMS2 and MSH6
        - Syntax: LIST (NOT_DONE, EXPRESSION, LOSS_OF_EXPRESSION) – SINGLE-VALUE
        - Units: N/A
        - Level: REQUIRED
    - Risk situation (only HNPCC)
        - 1..1 to person
        - Label: MM_RISK_SITUATION_HNPCC
        - Semantics:
            - Amsterdam criteria (Vasen HF, Watson P, Mecklin JP, Lynch HT (1999). "New clinical criteria for hereditary nonpolyposis

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

colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC". *Gastroenterology* **116** (6): 1453–6) OR Bethesda criteria (Umar A, Boland CR, Terdiman JP, et al. (2004). "Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability.". *J Natl Cancer Inst.* **96** (4): 261–268)

- Syntax: boolean (TRUE = risk situation)
- Units: N/A
- Level: OPTIONAL

- RAS mutation status
  - 1..1 to sample
  - Label: MM_RAS_MUTATION
  - Semantics:
    - TODO: link to ontology
  - Syntax: LIST (NOT_DONE, list of defined exons) – MULTI-VALUE
    - KRAS ex2 (codons 12 or 13), ex3 (codon 59 or 61), ex4 (codon 117 or 146); NRAS ex2 (codon 12 or 13), ex3 (codon 59 or 61), ex4 (codon 117 or 146)
    - For each of those, we need to know boolean yes/no (mutated/non-mutated); insertions/deletions/indels do not need to be considered
  - Units: N/A
  - Level: REQUIRED
- BRAF, PIC3CA. HER2 mutation status
  - 1..1 to sample
  - Semantics:
    - TODO: link to ontology
  - Syntax: LIST OF LISTS - one list of values per gene (NOT_DONE, MUTATED, NON_MUTATED) - default is NOT_DONE since people will mostly not have these done
    - List of relevant genes: BRAF, PIC3CA. HER2
  - Units: N/A
  - Level: OPTIONAL

- Histopathology part
  - 1..1 - to sample
  - TNM
    - Label: HIST_TNM
    - Semantics:

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

- existing guidelines
- TODO: UICC-TNM classification 7th edition

- Syntax: standard (TODO)
  - Three separate fields.
    - Primary tumor, takes the values (TX, T0, Tis, T1, T2, T3, T4a, T4b)
    - Regional lymph nodes, take the values (NX, N0, N1, N1a, N1b, N1c, N2, N2a, N2b)
    - Distant metastasis, takes the values (M0, M1, M1a, M1b)
- Units: N/A
- Level: REQUIRED

- UICC staging
  - Label: HIST_STAGING
  - Semantics:
    - standard – but changes over the time
    - TODO: link to ontology
  - Syntax: UICC (stage I to IV)
    - Two fields.
      - Stage, takes the values (0, I, IIA, IIB, IIIA, IIIB, IIIC, IVA, IVB)
      - Standard version, a list from 1st to 7th edition, or Not known
  - Units: n/a
  - Level: REQUIRED
  - Note: must include definition of the version of UICC standard used (can be implemented indirectly by providing date of determination of value)
- WHO grading
  - Label: HIST_GRADING
  - Semantics:
    - standard – but changes over the time
    - TODO: link to ontology
  - Syntax: WHO (G1 to G4)
  - Two fields.
    - Grade, takes the values (G1, G2, G3, G4)
    - Standard version, a list from 1st to 7th edition, or Not known
  - Units: n/a
  - Level: REQUIRED
  - Note: must include definition of the version of UICC standard used (can be implemented indirectly by providing date of determination of value)
- morphology
  - Label: HIST_MORPHOLOGY

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

        - Semantics:
            - standard – WHO
            - TODO: link to ontology
        - Syntax: standard (TODO)
        - Single field, takes the values (ADENOCARCINOMA, MUCINOUS_CARCINOMA, SIGNET-RING_CELL_CARCINOMA, MEDULLARY_CARCINOMA, HIGH-GRADE_NEUROENDOCRINE_CARCINOMA, LARGE_CELL_NEUROENDOCRINE_CARCINOMA, SMALL_CELL_NEUROENDOCRINE_CARCINOMA, SQUAMOUS_CELL_CARCINOMA, ADENOSQUAMOUS_CARCINOMA, MICROPAPILLARY_CARCINOMA, SERRATED_ADENOCARCINOMA, SPINDLE_CELL_CARCINOMA, MIXED_ADENONEUROENDOCRINE_CARCINOMA, UNDIFFERENTIATED_CARCINOMA, OTHER)
        - Units: n/a
        - Level: REQUIRED
    - Localization
        - Label: HIST_LOCALIZATION
        - Semantics:
            - standard – ICD-10
            - TODO: link to ontology
        - Syntax: standard (TODO) C18.1 to C18.7, C19, C20
        - Units: n/a
        - Level: REQUIRED
    - Metastasis
        - Label: HIST_METASTASIS
        - Semantics:
            - localization of metastasis
            - Based on ICD-10
            - TODO: link to ontology
        - Syntax: standard (TODO –)
        - Takes the values (NONE, PULMONARY, OSSEOUS, HEPATIC, BRAIN, LYMPH_NODES, BONE_MARROW, PLEURA, PERITONEUM, ADRENALS, SKIN, OTHERS)
        - Units: n/a
        - Level: REQUIRED
        - Note: yes/no is part of TNM
- Diagnostic exam (1..1 relation to patient)
    - Colonoscopy

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

- Label: DIAG_COLONOSCOPY
- Semantics:
  - whether colonoscopy was done
  - TODO: link to ontology
- Syntax: LIST (NOT_DONE, POSITIVE, NEGATIVE)
- Units: n/a
- Level: REQUIRED
- Note: possible relation/interaction with the localization in histopathological diagnosis

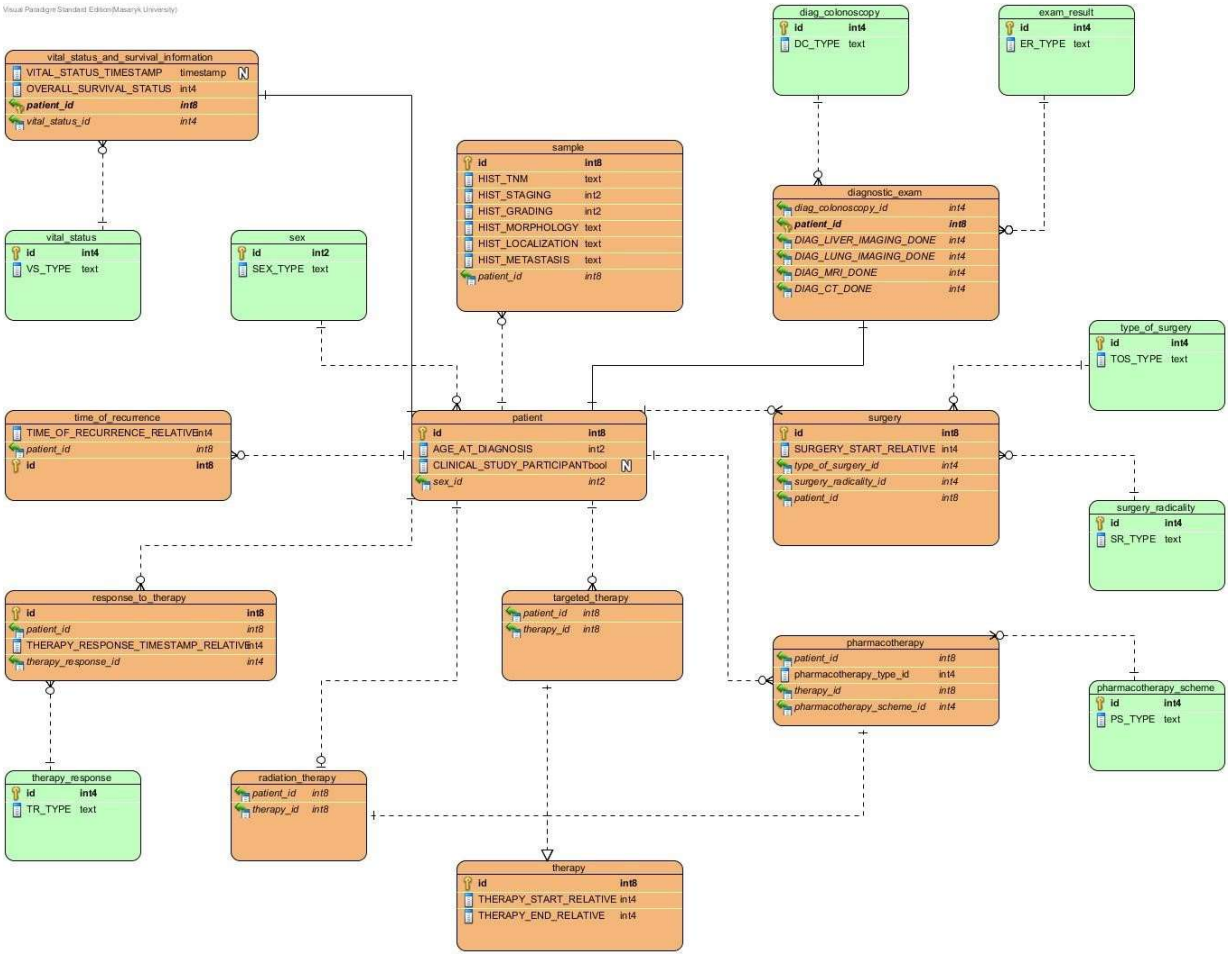
o Array of diagnostic methods (liver imaging, lung imaging, MRI, CT)

- Label: DIAG_X_DONE
  X l1 {LIVER_IMAGING, LUNG_IMAGING, MRI, CT}
- Semantics:
  - whether given diagnostics was done
  - TODO: links to ontologies
- Syntax: list (NOT_DONE, DONE_DATA_AVAILABLE, DONE_DATA_NOT_AVAILABLE, UNKNOWN)
- Units: n/a
- Level: REQUIRED

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

ENTITY RELATION MODEL

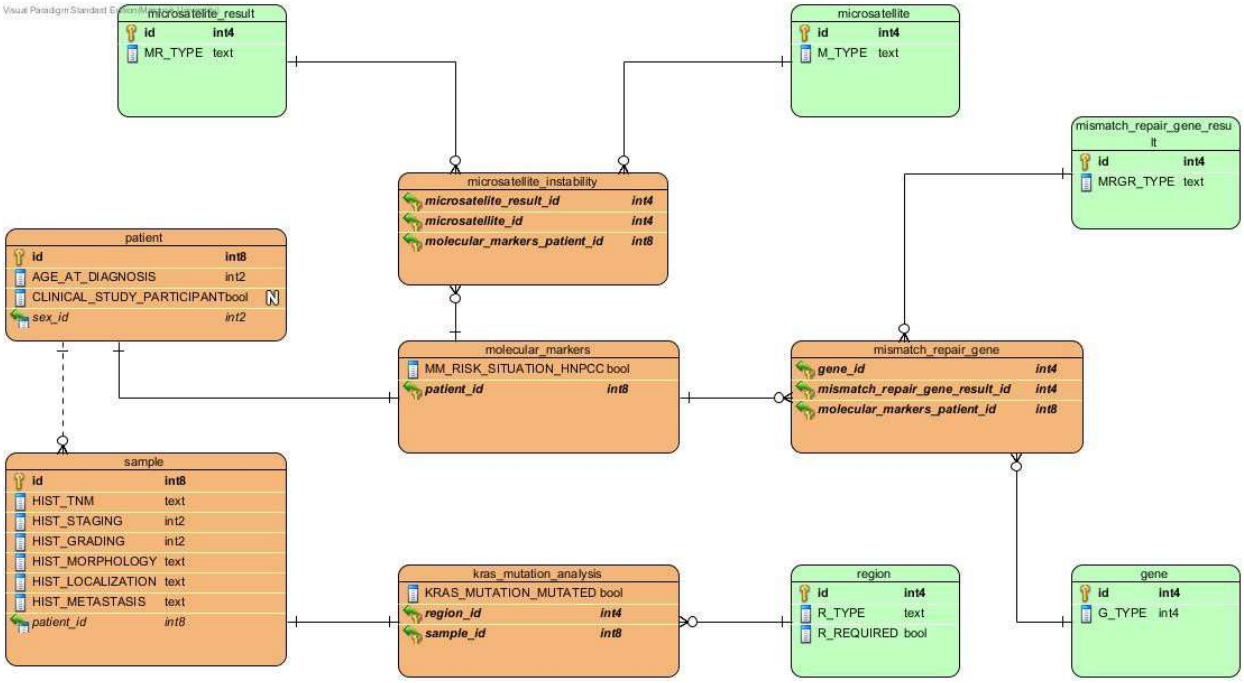
Patient section



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

ADOPT BBMRI-ERIC gateway for health

Molecular markers section



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

ADOPT BBMRI-ERIC gateway for health

RESOURCES TO BE REVIEWED

- http://www.nejm.org/doi/full/10.1056/NEJMp1401444
- CDISC
- http://openehr.org/
- http://www.iccr-cancer.org/datasets
- RedCap questionnaires related to colon cancer
- National standards
    - Czech NOR
    - German standards
    - Dutch registry

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