Coordinated Research Infrastructures Building Enduring Life-science services
- CORBEL -

Deliverable D3.10
Report on final assessment and generalization of integration platform

WP3 – Community-driven cross-infrastructure joint research - Medical

Lead Beneficiary: ECRIN-ERIC
WP leader: Jacques Demotes (ECRIN-ERIC)
Contributing partner(s): Lygature (EATRIS), ErasmusMC, NKI (EATRIS), Istituto Ortopedico Rizzoli (IOR - EATRIS), Università Degli Studi di Torino (UNITO), University Medical Centre Utrecht (UMCU)

Contractual delivery date: 31 January 2020
Actual delivery date: 27 January 2020

Authors of this deliverable: Jan-Willem Boiten (Lygature), Silvio Aime (UNITO), Mariska Bierkens (NKI), Walter Dastrù (UNITO), Mattias Hansson (Erasmus MC), Stefan Klein (Erasmus MC), Dario Longo (UNITO - IBB CNR), Sjaak Peelen (Lygature), Sara Zullino (UNITO), Luca Sangiorgi (IOR), Marina Mordenti (IOR), Manila Boarini (IOR), Paco Welsing (UMCU)

Grant agreement no. 654248
Horizon 2020
H2020-INFRADEV-1-2014
Type of action: RIA

Content

Executive Summary

Task 3.4 in CORBEL ("Data integration and management services for image-driven and genomics-driven biomarker studies") was designed to develop a common IT framework to support data handling and analysis for both clinical and preclinical biomarker research, across disease areas. The effectiveness of the selected approaches has been demonstrated using national and international research projects in this domain. It was investigated whether the data integration and imaging solutions implemented for the selected use cases are not only operational and used, but can also serve as best practices within the framework of shared services between the (ESFRI) Biological and Medical Research Infrastructures in the medical domain. This has been demonstrated with four independent use cases across the spectrum from preclinical to clinical biomarker research:

1) In the preclinical research use case tools were developed for easily extracting, importing, archiving preclinical studies based on magnetic resonance imaging (MRI) from several vendors, including tools for automated image processing. The herein developed workflow will be made available to the wider preclinical research community, hence allowing a simplified exchange and (re)use of image datasets among preclinical imaging centers.
2) In the PROOF [1] use case (osteoarthritis), a generic infrastructure was developed for integrative analysis of medical imaging, genetics and clinical data, supporting imaging-genetics association studies and advanced radiomics analyses using multivariate machine-learning to develop novel predictive markers. Best practices are given.
3) In the IMI APPROACH [2] use case (osteoarthritis), a generic infrastructure was implemented largely shared with the other two clinical use cases. This integration solution is operational and successfully used in the daily project workflow. The solution regarding the data processing and loading is set up to be generic and the final approach taken can be used as a best practice for future projects, which is further detailed in the recommendations towards the end of the present document.
4) For Cancer Core Europe [3], a collaboration project managed between a number of leading European cancer centres, a data sharing taskforce has been set up to implement the data integration solutions. Within this project a use case revolving around patients with non-breast related tumours in conjunction with a BRCA1 or BRCA2 mutation, was selected to execute with support of CORBEL. The implementation of the integration solutions are discussed and conclusions and recommendations shared.

The core of the CORBEL common IT framework for genomics and imaging studies in biomarker research consists of the tranSMART (data integration and browsing platform), cBioportal (patient-level genomics platform) and XNAT (image archive platform) solutions. The present document describes how this infrastructure facilitates multi-modal biomarker research. The four use cases highlight what is needed for multimodal biomarker research, such as data capture, data curation pipelines, standardization, transformation, integration, and analysis. We conclude with recommendations regarding practical use of the common IT framework for future projects.

Project objectives

With this deliverable, the project has contributed to the following objective:

a) Establish generic data integration services for image-driven and/or genomics-driven translational studies (e.g. biomarker discovery in heterogeneous diseases) embedding cross-RI services.

Detailed report on the deliverable

Background

WP3 coordinates the activities of a series of tasks (WP3.1 to WP3.5) devoted to the development of common tools - each involving at least two of the ESFRI biomedical research infrastructures - to foster integration and interoperability of research infrastructures supporting the development of innovative prevention, diagnostic or treatment solutions.

Biomarker research projects have to deal with a broad diversity of data (phenotypic data, clinical outcomes, images, genomics, biosamples, etc.), each requiring different tools and methods. Individual hospitals often lack the number of patients and the spectrum of techniques required for efficient biomarker research or precision medicine approaches in general, therefore projects are increasingly supported by large international consortia and public-private partnerships. However, even the largest biomarker research programs face difficulties in timely managing their data sets, for clinical as well as preclinical biomarker and personalised medicine research, due to the lack of multimodal data infrastructure for standardized storage, management and processing. Yet because between translational research studies there is such a common, similar, design, the same issues in handling these data types are encountered time and time again, making it possible to offer solutions to these issues in a structural, standardized manner. CORBEL task 3.4 was designed to establish a common IT framework to support data handling and analysis for both preclinical and clinical biomarker research, across different disease areas. This could be successfully achieved by bundling the expertise and services of the biomedical ESFRIs, most notably EATRIS, EuroBioImaging, ELIXIR and BBMRI, providing solutions for biomarker development projects. The effectiveness of the selected approaches was demonstrated using international research projects in this domain.

The selected IT framework consists of a number of key solutions briefly introduced below:

- tranSMART: the central data integration platform;
- cBioPortal: patient-centric and gene-centric viewing of genomics data;
- XNAT: the central clinical imaging archive.

The current report summarizes the findings from the use cases and provides recommendations for future usage of the selected technologies in similar research projects.

Data integration platform: tranSMART

tranSMART is an open-source data integration, sharing, and analysis platform for clinical and translational research and constitutes the core of the IT infrastructure in the CORBEL clinical use cases for biomarker research. The platform allows users to search, view, and analyze ‘final’ or ‘processed’ data through a web interface, thereby allowing easy access to explore such data from multiple domains at a study, or cohort level (Figure 1). It also enables scientists to develop and refine hypotheses by investigating correlations between genetic, phenotypic and clinical data. Furthermore, domain experts or bioinformaticians can download the available data if they wish to do more complicated analyses. Finally, the possibility to add metadata (including hyperlinks) allows users to find other, related data external to the tranSMART data-integration platform, such as raw or pre-processed Next Generation Sequencing data.

Data acquisition, quality control and data processing pipelines are an integral part of the technology platforms for biomarker development projects (represented by our clinical use cases). The output of these specific data pipelines has been uploaded into tranSMART using well-established standards for molecular data formats (DNA, RNA, protein) and imaging data, where possible.

The open-source community of tranSMART is supported by a Foundation: the I2B2 tranSMART Foundation[1]. At the time of writing this report, multiple parallel versions of the software were in use in the community with unfortunately some divergent functionalities. The use cases supported by CORBEL have generally made use of tranSMART version 16.2.

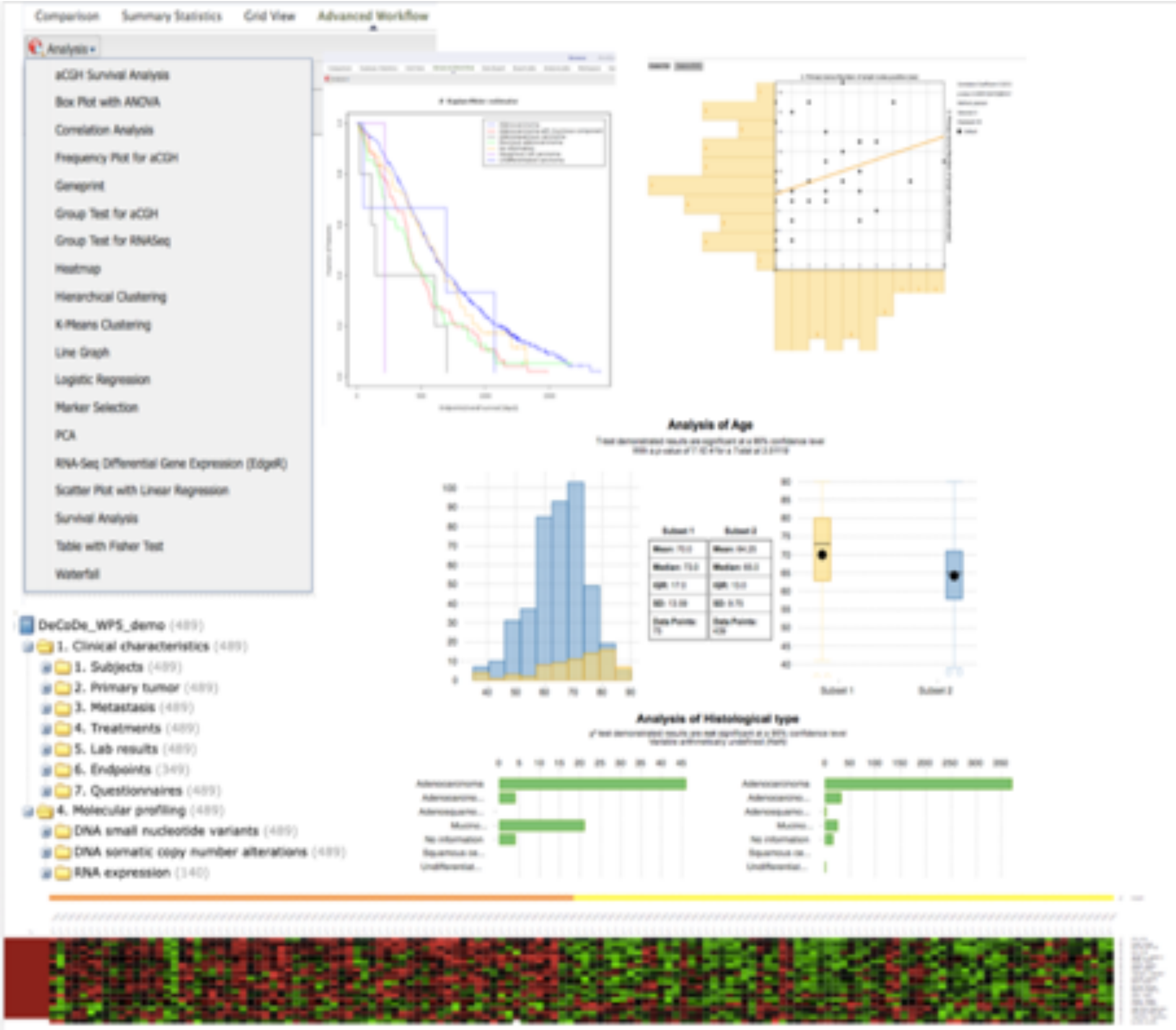


Figure 1: A compilation of several analyses of the public "DeCoDe_WP5_demo" study in tranSMART, with in the top left corner: different advanced analyses available within the application, bottom left corner: an example of a modelled study, giving a quick overview of the data available for that study, right: Kaplan-Meier Overall Survival Plot, correlation analysis, t-test, X2 and a gene-based heatmap.

[1] https://transmartfoundation.org/

With different types of data available within tranSMART it is now possible to examine existing data easily in an exploratory manner, thereby allowing researchers to optimally use their existing data.

Cancer genomics portal: cBioPortal

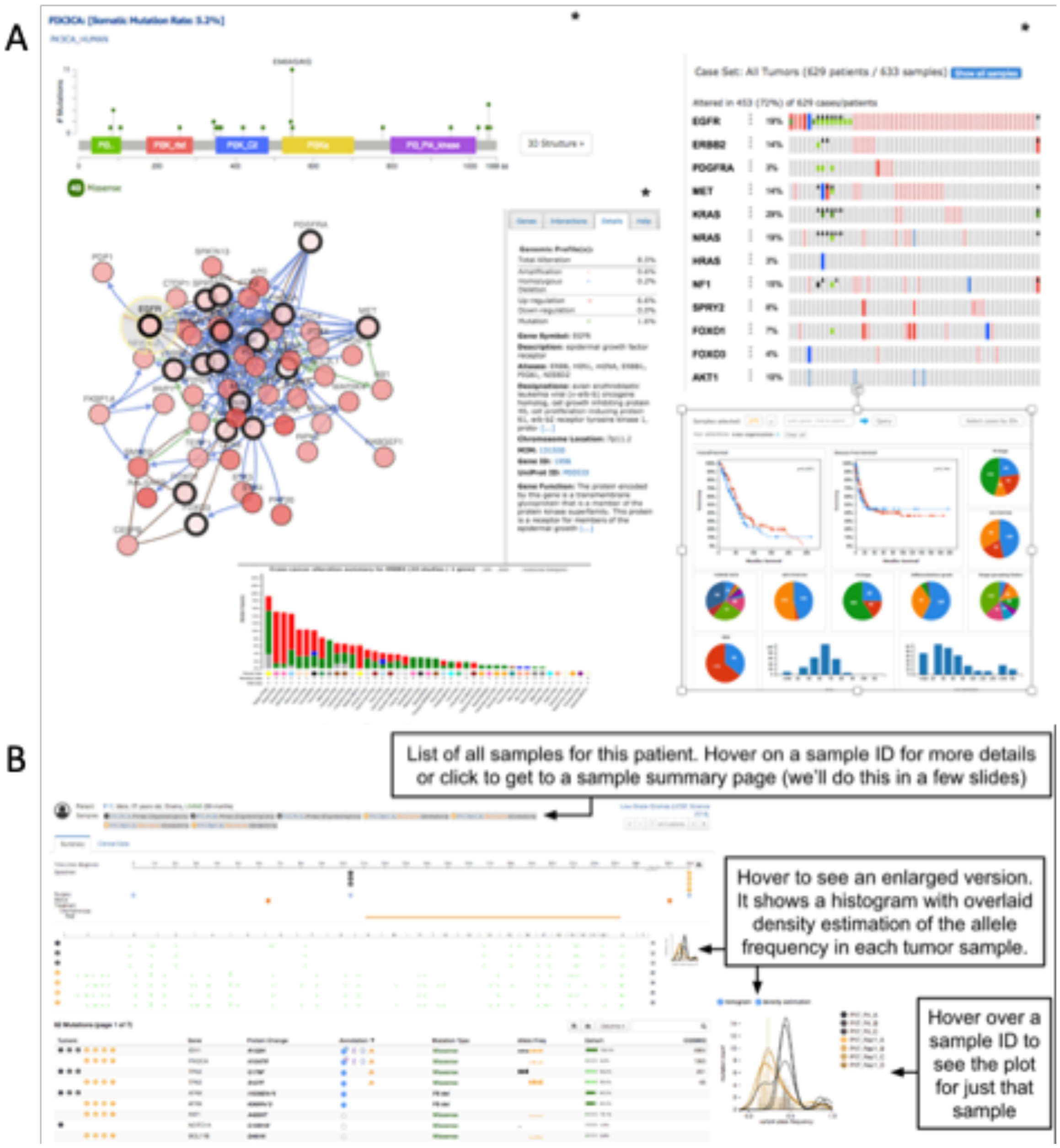
The cBio Cancer Genomics Portal[2], is an open-source resource for interactive exploration of multidimensional cancer genomics data sets, currently providing access to data from more than 82,875 tumor samples from already 273 public cancer studies. The cBio Cancer Genomics Portal significantly lowers the barriers between complex genomic data and (cancer) researchers who want rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large-scale cancer genomics projects, and empowers researchers to translate these rich data sets into biological insights and clinical applications [4]. It now also serves oncologists in their use and interpretation of clinical sequencing data from cancer patients enabling precision oncology.

In cBioPortal it is also possible to examine data within one or more studies, and to zoom in from a study's summary to a subject's individual case record, where case record data can be viewed on a time-axis as well (Figure 2). Furthermore, there is an embedding functionality within cBioPortal, which allows for a study's data owner to present images of a certain case record, for instance, scanned pathology images. With these additional functionalities, cBioPortal serves as a data-integration platform with complementary functionality to tranSMART.

The open-source community is actively coordinated from Memorial Sloan Kettering cancer hospital with increasing participation from centers across the world. The cBioPortal software is also actively used by leading research consortia, including the TCGA project [6][3].

[2] http://cbioportal.org

[3] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga


A
PIK3CA: [Somatic Mutation Rate: 5.2%]
PIK3CA_HUMAN
Missense
3D Structure
Case Set: All Tumors (629 patients / 633 samples)
Altered in 453 (72%) of 629 cases/patients
EGFR
ERBB2
PDGFRA
MET
KRAS
NRAS
HRAS
NF1
SPRY2
FOXO1
FOXO3
AKT1
EGFR
PDGFRA
MET
B

List of all samples for this patient. Hover on a sample ID for more details or click to get to a sample summary page (we'll do this in a few slides)

Hover to see an enlarged version. It shows a histogram with overlaid density estimation of the allele frequency in each tumor sample.

Hover over a sample ID to see the plot for just that sample

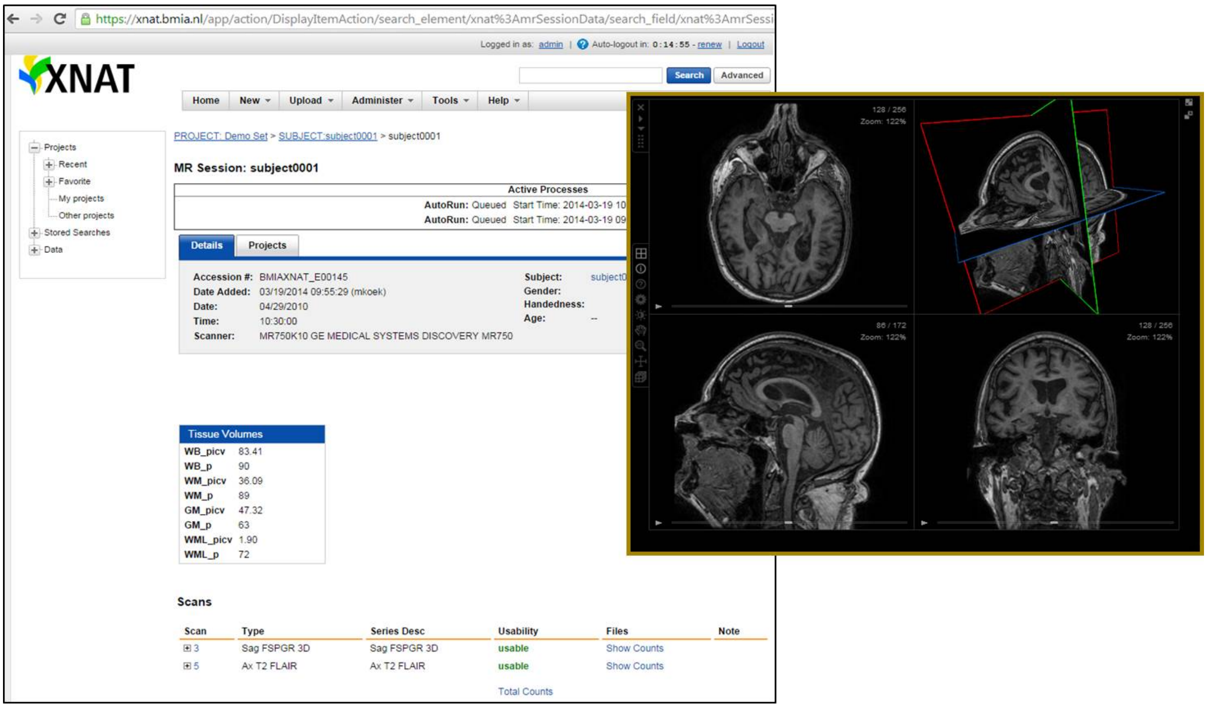
Figure 2: Just a few of the different view and query possibilities within cBioPortal, allowing for a gene-centric exploration of available research data. (A) Top left: mutation 'lollipop' figure highlighting the mutation domains within one gene for all selected cases; middle left: biological network data, retrieved from Pathway Commons, displaying the queried genes and their alteration status within the network; bottom left: overview of alteration types for one gene, across multiple studies; top right: the oncoprint (geneprint), displaying for a multitude of queried genes the alteration status, marked with different colours and symbols; bottom right: summary overview page of study data, including Kaplan-Meiers for Overall and Progression Free Survival, as well as pie-charts and histograms for categorical and numerical data (screenshots taken from colorectal cancer study in cBioPortal.org [5]). (B): visualisation of sample and longitudinal data for one patient (screenshots from the cBioPortal online tutorial #3: Patient View[4]).

[4] https://www.cbioportal.org/tutorials

Imaging platform: XNAT

XNAT[5] [7] is an open-source imaging informatics platform developed by the Neuroinformatics Research Group at Washington University. XNAT was originally developed in the Buckner Lab at Washington University, now at Harvard University. It facilitates common management, productivity, and quality assurance tasks for imaging and associated data. Thanks to its extensibility, XNAT can be used to support a wide range of imaging-based projects.

XNAT enables data access via a website (manual upload and download), via the DICOM protocol, and via an application programming interface (API). Furthermore, XNAT stores not only the images, but also image-derived information, such as annotations and processed versions of the images. It is therefore especially of interest for the more advanced, technically oriented researchers, and for large studies which require automated image analysis. Projects are structured in a standard way according to the hierarchy Project->Subject->Experiment->Scan. Access rights for each user can be configured in detail (read only, read-edit, read-edit-delete; depending on data type and project); such functionality is crucial when dealing with sensitive personal data like medical images. A screenshot of the interface is shown in Figure 3.



PROJECT: Demo Set > SUBJECT:subject0001 > subject0001

MR Session: subject0001

Active Processes

AutoRun: Queued Start Time: 2014-03-19 10

AutoRun: Queued Start Time: 2014-03-19 09

Accession #: BMIAXNAT_E00145
Date Added: 03/19/2014 09:55:29 (mkoek)
Date: 04/29/2010
Time: 10:30:00
Scanner: MR750K10 GE MEDICAL SYSTEMS DISCOVERY MR750

Subject: subject0
Gender:
Handedness:
Age: --

Tissue Volumes	
WB_picv	83.41
WB_p	90
WM_picv	36.09
WM_p	89
GM_picv	47.32
GM_p	63
WML_picv	1.90
WML_p	72

Scans

Scan	Type	Series Desc	Usability	Files	Note
⊞ 3	Sag FSPGR 3D	Sag FSPGR 3D	usable	Show Counts	
⊞ 5	Ax T2 FLAIR	Ax T2 FLAIR	usable	Show Counts	
			Total Counts		

Figure 3: A screenshot of the web interface of XNAT showing data for an example scan session. Both brain MRI DICOM images and derived brain tissue volume measurements are stored in this example. XNAT also includes a built-in javascript-based image viewer shown on the right. For more advanced visualization functionality, XNAT can be connected to external viewers, either using the DICOM interface or the (HTTP-based RESTful) API.

Description of Work

The findings on whether the selected solutions can work as generic data integration services are reported in this document. Based on the use cases, the selected tools can act as a common IT

[5] https://www.xnat.org/

framework to support data handling and analysis for both clinical and preclinical biomarker research, across the disease areas. The challenges however, lie in the ETL (Extract, Transform, Load) processes as well as in standardization of the metadata describing the datasets to be uploaded. Basically, this is a prerequisite for practical usage of any IT framework in this domain. Therefore, a set of procedures is needed to make the data loading effective, reproducible and broadly accessible by any user. Such procedures consist of technical solutions (scripts, programs, etc.), as well as documentation describing prerequisites and best practices. The procedures used by the four use cases are described in the paragraphs below. Furthermore, it is described what the challenges are concerning the standardization of (meta)data when integrating data from different sources, and recommendations for future projects are given.

Preclinical use case

Small animal imaging facilities are highly specialized centers that provide the research community access to cutting-edge imaging technologies. These centers therefore have to deal with the complexity and the variety of preclinical trial data sets. Moreover, imaging data analysis requires a multidisciplinary effort, in terms of data management and processing. This has also prompted increasing interest in the development of data-driven models based on computational approaches and image processing algorithms. The difficulties to overcome in such projects depend on the complexity of the processing and on the scarcity of standard tools for sharing and processing medical images across imaging centers and/or acquired with different imaging equipment.

Our aim is to overcome these limitations through the integration of an open-source archiving platform with customizable tools for automated image processing. The workflow is available to the preclinical research community, hence allowing a simplified exchange and reuse of image datasets between preclinical imaging facilities.

We have developed Python/Matlab-based tools for exporting, processing and archiving preclinical images using the built-in Pydicom library [8]. These tools interface with the remotely accessible XNAT [7] database, a widely used open-source platform for managing, sharing and processing medical imaging DICOM data, via XNAT Python clients XNATpy[6] and PyXNAT[7] [9]. XNAT natively supports multiple imaging modalities, such as MRI, PET, CT, and US.

Since preclinical imaging instrumentation adopts a proprietary format, a tool for converting raw images to the DICOM standard has been developed. We demonstrate the fit of this tool to our workflow with MRI images and BRUKER instrumentation as a use case scenario. The workflow is based on the following steps (Figure 4):

- uploading imaging datasets to XNAT in DICOM standard acquired through several instrumentations and modalities (1);
- a Bruker ParaVision to DICOM format converter to import MRI images to XNAT (2);
- an XNAT image processing pipeline accepting DICOM files as input (3) to produce parametric images by calling custom image-analysis scripts.

[6] https://xnat.readthedocs.io

[7] https://pyxnat.github.io/pyxnat/

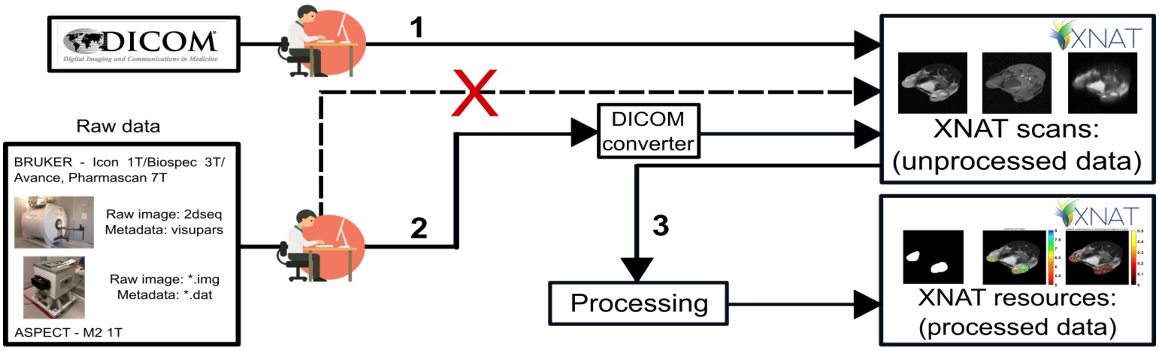
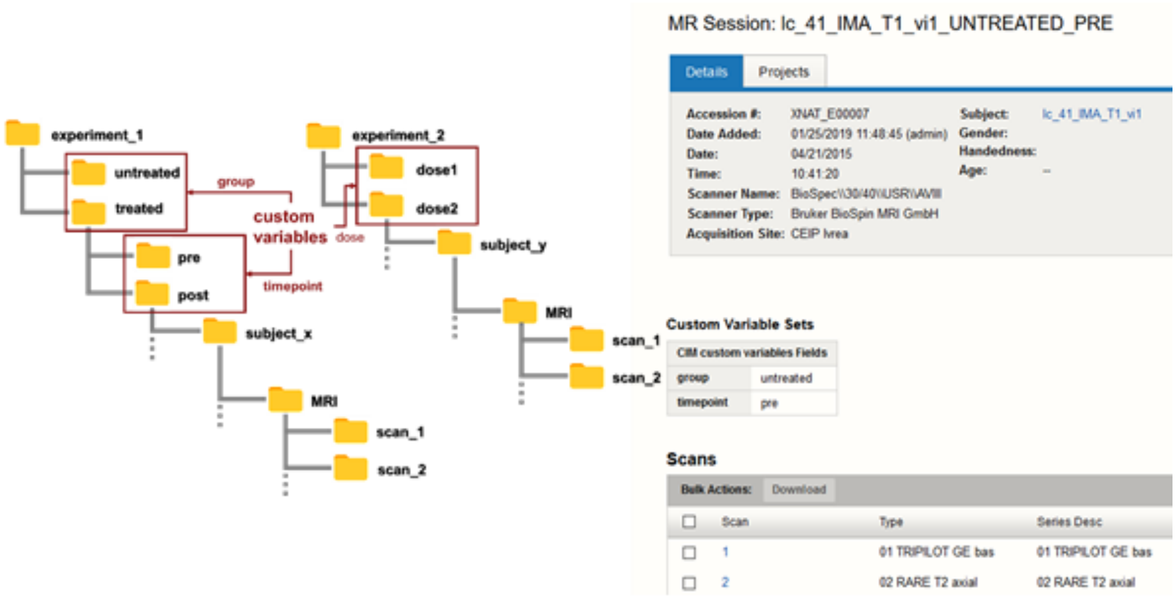

DICOM
Digital Imaging and Communications in Medicine
Raw data
BRUKER - Icon 1T/Biospec 3T/ Avance, Pharmascan 7T
Raw image: 2dseq
Metadata: visupars
Raw image: *.img
Metadata: *.dat
ASPECT - M2 1T
1
2
3
DICOM converter
Processing
XNAT
XNAT scans: (unprocessed data)
XNAT
XNAT resources: (processed data)

Figure 4: Schematic workflow of image archiving and processing.

In addition, to deal with the new MRI-based technique of Chemical Exchange Saturation Transfer (CEST), private DICOM tags specifically devoted to this modality have been introduced. To avoid conflicts of private tags from other implementers, we reserved a section in the standard DICOM dictionary for this specific use. The dedicated DICOM dictionary has been extensively tested on MRI-CEST data, allowing users to efficiently manage and label a large amount of imaging data sets. Upon conversion to DICOM, the image data set can be uploaded to XNAT. To manage different experimental protocols, XNAT data types have been extended using custom variables. These user-defined variables refer to the treatment administration (treated/untreated groups), different timepoints, and drug doses (Figure 5).



MR Session: lc_41_IMA_T1_vi1_UNTREATED_PRE

Accession #: XNAT_E00007
Date Added: 01/25/2019 11:48:45 (admin)
Date: 04/21/2015
Time: 10:41:20
Scanner Name: BioSpec\\30/40\\USR\\AVIII
Scanner Type: Bruker BioSpin MRI GmbH
Acquisition Site: CEIP Ivrea
Subject: lc_41_IMA_T1_vi1
Gender:
Handedness:
Age: --

Custom Variable Sets

CIM custom variables Fields	
group	untreated
timepoint	pre

Scans

Bulk Actions: Download

□	Scan	Type	Series Desc
□	1	01 TRIPILOT GE bas	01 TRIPILOT GE bas
□	2	02 RARE T2 axial	02 RARE T2 axial

Figure 5: Schema of the data hierarchy with custom variables (left panel); screenshot of the MRI session webpage on the XNAT database (right panel).

XNAT comes with a pipeline engine that can run external applications and shell scripts. Pipelines can be executed on project level, either by selecting the scan of interest or letting the application look automatically for the DICOM tag that specifies the acquisition protocol. In Figure 6, an example schema for processing Diffusion Weighted Imaging (DWI) acquisitions is presented.

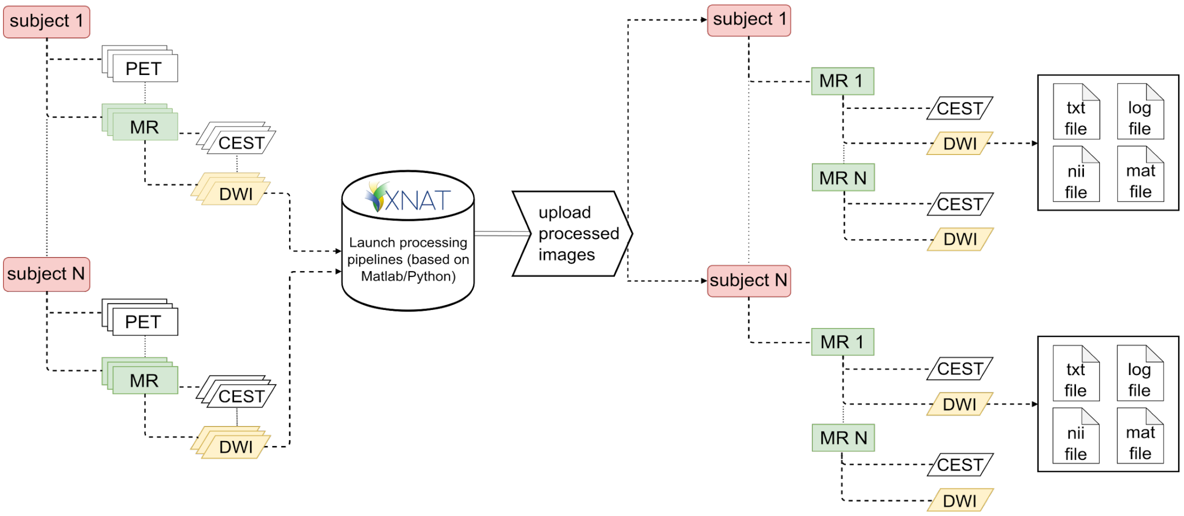


Figure 6: Schema of an XNAT pipeline which retrieves, downloads and processes all the DWI scans in a project. The output files (text file, log file, NIfTI file and Matlab workspace) are then uploaded back into XNAT under the corresponding subject, experiment and scan.

Processed data and other output files are uploaded back into XNAT in the resource folder available at each level of the XNAT data structure and accessible by the user through a Manage Files console. The resource folder can contain several subfolders, each of them populated with binary masks, MATLAB files, log files and many others. Representative workflow examples have been presented as conference abstracts [10] showing use cases dealing with MRI-based preclinical images.

Conclusions

A user-friendly, customizable workflow has been developed to 1) convert MRI proprietary raw data to medical image standard format (DICOM) and 2) to upload whole image datasets to an XNAT-based archive system. The workflow can manage several imaging modalities (MRI, PET, CT and US) and different types of preclinical investigation protocols using newly developed XNAT custom variables. An instance of the XNAT platform containing the developed tools is available and accessible at the following link[8].

Besides facilitating storage of raw, unprocessed imaging data, analysis tools have been implemented for automated image processing and storage of image-derived data that result from quantitative analysis. We believe that such a workflow could be of interest for preclinical imaging centers, thus allowing the scientific community to efficiently store, process and share biomedical imaging data.

PROOF use case

Background

Precision medicine requires deep understanding of the relationship between genes and phenotypes. In the field of genetics, genome wide or candidate gene association studies are typically performed using readily available but imprecise clinical phenotypes as the correlate of interest. While such analyses have led to important discoveries, there is an emerging interest in genetic association studies with deep phenotypes [11], which represent quantitative and objective biological markers instead of qualitative and often subjective clinical assessments. In-vivo and clinical medical imaging modalities, such as MRI, CT, and PET, provide rich information about tissue properties, patient

[8] http://www.cim.unito.it/website/research/research_xnat.php

anatomy and pathology. Medical imaging potentially therefore stands at the basis of many deep phenotypes. Visual assessment and semi-quantitative scoring of images already reveals much information, but to truly exploit the rich information that the images provide, quantitative imaging biomarkers should be used as deep phenotypes. Examples are the volume of the hippocampus as an MRI-derived biomarker for Alzheimer's disease or the volume of the knee cartilage as an MRI-derived biomarker for knee osteoarthritis, but also more advanced computational measures based on radiomics and machine learning approaches [12]. To enable large-scale studies and ensure reproducibility, methods and software are needed to automate the computation of quantitative imaging biomarkers, and to facilitate their integration with high-dimensional genetic data to enable large-scale imaging-genetics association studies, correlating genes to quantitative imaging biomarkers. To this end, a generic infrastructure was developed for integrative analysis of medical imaging, genetics and clinical data, supporting imaging-genetics association studies and advanced radiomics analyses using multivariate machine-learning to develop novel predictive markers.

As a use case study guiding our developments, we used data from a study on the PRevention of knee Osteoarthritis in Overweight Females – PROOF [1]. The PROOF study is a randomized-controlled trial, investigating the effect of preventive strategies on the development of osteoarthritis (OA) in a high-risk population. 407 overweight females were included, with follow-up of 6.5 years. Besides clinical and metabolic measures, morphological MRI and radiographs for both knees, and genome-wide genotyping are available. This rich data collection, with multiple samples per subject with longitudinal data collection, makes it a perfect test case highly representative for many future studies that require multimodal data integration, including but not limited to other studies on osteoarthritis such as IMI APPROACH (use case 3 in this CORBEL task; see below for more details).

Results

As described in detail in previous deliverable and milestone reports, we have developed a generic infrastructure for integrative analysis of medical imaging, genetics and clinical data, supporting both imaging-genetics association studies and advanced radiomics analyses. The infrastructure has been designed in a modular way and built on top of existing open-source software solutions, which have been made interoperable. The flowchart in Figure 7 provides an overview of the infrastructure.

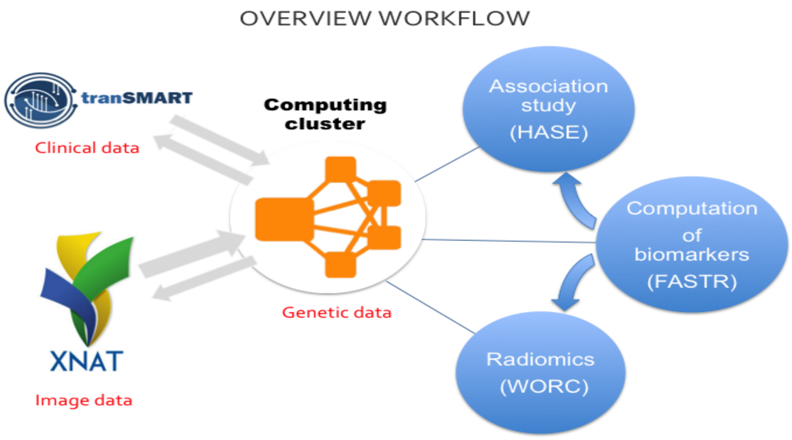


Figure 7: Flowchart of the modular infrastructure implemented in the context of the PROOF use case, enabling integrative analysis of medical imaging, genetics and clinical data, supporting computation of imaging biomarkers (using FASTR), imaging-genetics association studies (using HASE), and

advanced radiomics analyses (using WORC). Clinical data is stored in tranSMART, imaging data is stored in XNAT, and genetic data is stored in a compressed format on the compute cluster (within the firewall of our institute, with read access for selected users only) to ensure efficient data access by the compute nodes. All communication with XNAT is automated, for which we use a Python toolbox called XnatPy, which wraps and simplifies the RESTful API of the XNAT web service. Subsets of clinical data can be easily exported from tranSMART and stored on the compute cluster for analysis purposes.

The infrastructure is fully operational as demonstrated previously in Milestone Report MS21, by two pilot experiments using data from the PROOF study, illustrating 1) imaging-genetics association analysis by simple regression between meniscus volume measurements (automatically computed from the MRI scan) and single nucleotide polymorphisms (SNPs), and 2) radiomics analysis to predict clinical outcomes (Kellgren-Lawrence grade) using multivariate machine learning based on hundreds of computational image features derived from MRI.

After these initial pilot experiments, we have used (elements of) this generic and modular infrastructure in follow-up research projects, leading to 1) a conference abstract [13] and a submitted journal manuscript analysing the correlation of meniscus volume changes with several clinical OA outcomes in the PROOF study, 2) a conference abstract [14] presenting a radiomics study using knee MRI data of a large population-based cohort (the Rotterdam Scan Study [15]), to assess the ability of radiomic features to discriminate subjects with and without medial tibial osteophytes. These successful studies demonstrate the reusability of the infrastructure.

Conclusions

- A modular infrastructure for integrative analysis of medical imaging, genetics and clinical data was proposed. The infrastructure combines open-source solutions for data storage (XNAT, tranSMART), computation of imaging biomarkers (FASTR), imaging-genetics association analysis (HASE), and advanced radiomics analyses (WORC).
- Initial pilot experiments and later research projects (on PROOF data, but also on another data collection) demonstrate the usability of the infrastructure.

APPROACH use case

The IMI APPROACH project aims to gain a better understanding of osteoarthritis (OA) disease phenotypes and acceptance of a shared guideline to classify or stratify patients. This stratification approach will provide clear and specific phenotype-directed protocols for disease modifying OA drug trials enabling the targeting of subgroups with OA that have uniform disease characteristics, thereby increasing the chances of success. One of the project's objectives to achieve such stratification approach using data from many different sources, is "Implement and establish a new, integrated and comprehensive database platform of existing data from partners that will be extended with newly collected longitudinal data, incorporating novel high-quality biomarkers".

In order to meet this goal from the APPROACH project, retrospective cohort data from multiple cohorts had to be uploaded into tranSMART. This complicated integration effort gave rise to multiple challenges to be overcome. While processing the individual data files for upload into tranSMART, it became apparent that the quality and standardization of the metadata of the different cohort data sets varies significantly. Some cohorts have an extensive unambiguous codebook, whereas others only have a minimal description of the parameters stored. It was also observed that, even in some cases where sufficient metadata were provided, these data were scattered over different files. These files were geared towards human readability, but were not suited for automated extraction of the required metadata. Furthermore, there was hardly any standardization between the different osteoarthritis cohorts with regards to the naming and description of the parameters used. As a result, the relevant metadata had to be extracted mostly manually, which is not only a time-consuming process, but also rather error-prone.

Because tranSMART requires certain data formats for importing, the original source data files had to be pre-processed before they could be loaded into tranSMART via the tranSMART-batch pipeline. Details on the pre-processing in the APPROACH project are described in more detail below. Extended documentation on data formats accepted for input to the tranSMART-batch pipeline can be found in the **Corbel deliverable report Deliverable D3.9 Robust upload procedures final and here** [16][9]. This reference both describes upload of low and high dimensional data. However, for the APPROACH project only upload of low-dimensional data is applicable. These low dimensional data include clinical data, various image analysis results, like Kellgren-Lawrence [17] scores, gait analysis data, optical hand scan data and biomarker data. These are 'simple' data, for which there is one numerical or categorical observation of a concept for each subject (e.g. subject gender (Female/Male) or a specific disease score). The APPROACH data are, however, longitudinal as several of the observations are repeated over time (at the various clinical visits or follow ups). For the tranSMART version (16.2) used, longitudinal data are in principle not supported and reformatting of the data is needed to store the data in such a manner that the longitudinal aspect in included. More recent releases of tranSMART (17.1) or the related I2B2 solution [18][10] are able to handle longitudinal data. So it is advised for new projects that work with longitudinal data to consider tranSMART 17.x or I2B2, although the support for molecular biology data is more limited in these tools.

ETL – APPROACH pre-processing pipeline

The APPROACH pre-processing pipeline is set up by a number of (Perl) scripts and a codebook (Figure 8). This codebook indicates the mapping between the original parameter and how (and if) this parameter should be visible in the tranSMART hierarchical tree. This codebook also provides the necessary metadata annotating the primary study data, such as:

- description of the parameter
- expected values and limits
- units
- whether a parameter is numeric or categorical

The codebook file should be a delimited text file (e.g. csv or Excel file) containing a table with the columns as indicated in Table 1. The longitudinal parameter information is added to the tranSMART codebook by one of the processing scripts. Besides using the codebook for organising the data in tranSMART, the codebook also provides metadata information which is uploaded into tranSMART.

[9] https://github.com/thehyve/transmart-batch/tree/master/docs

[10] https://www.i2b2.org/

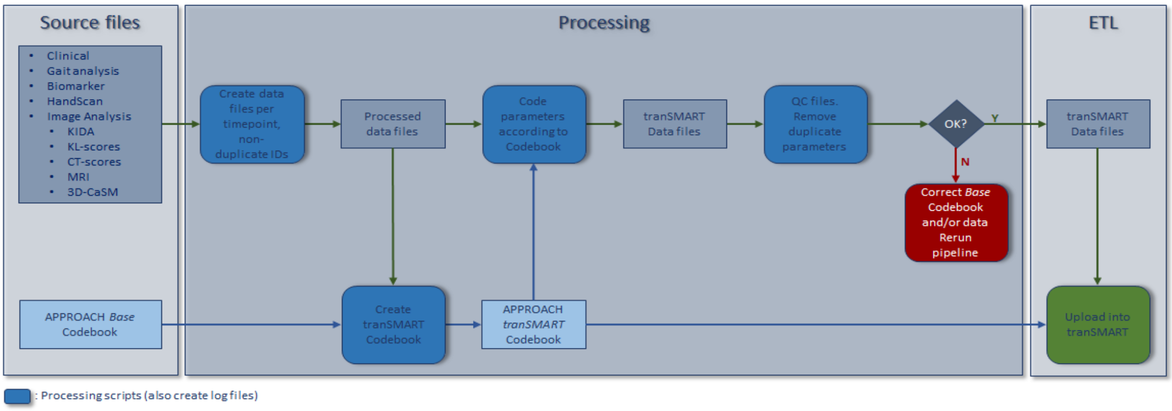

Source files
Clinical
Gait analysis
Biomarker
HandScan
Image Analysis
KIDA
KL-scores
CT-scores
MRI
3D-CaSM
APPROACH Base Codebook
Processing
Create data files per timepoint, non-duplicate IDs
Processed data files
Code parameters according to Codebook
tranSMART Data files
QC files. Remove duplicate parameters
OK?
Y
N
Correct Base Codebook and/or data Rerun pipeline
Create tranSMART Codebook
APPROACH tranSMART Codebook
ETL
tranSMART Data files
Upload into tranSMART
: Processing scripts (also create log files)

Figure 8: APPROACH tranSMART pre-processing pipeline

Table 1. Codebook data

Column name	Description
Category 1	Main parameter category
Category 2	Parameter sub-category (optional)
Category 3	Parameter sub-sub-category (optional)
Category 4	Parameter sub-sub-sub-category (indicating Left or Right) (optional)
Baseparameter	Parameter name, time point independent
Parameter	Parameter name including the timepoint prefix
Timepoint	Coded visit indicator (prefix of Parameter (e.g. M006)
Visit	Textual description of Timepoint, describing when a parameter value was determined (e.g. Month 06).
Description	Short description of the (Base)parameter
Categorical	Description whether the parameter is categorical or not (Y/N). If Y, this forces numerical values as being categorical in tranSMART.
Values	Possible values the parameter is allowed to have. Coded information is indicated using a "=" (*e.g.* 0=No;1=Yes).
Parameter@Source	Name of a parameter in the source data file.
Source	System from which the source data originate.

Figure 9 shows a screenshot of the tranSMART hierarchical tree and illustrates how the codebook columns are used to set up this tranSMART tree.

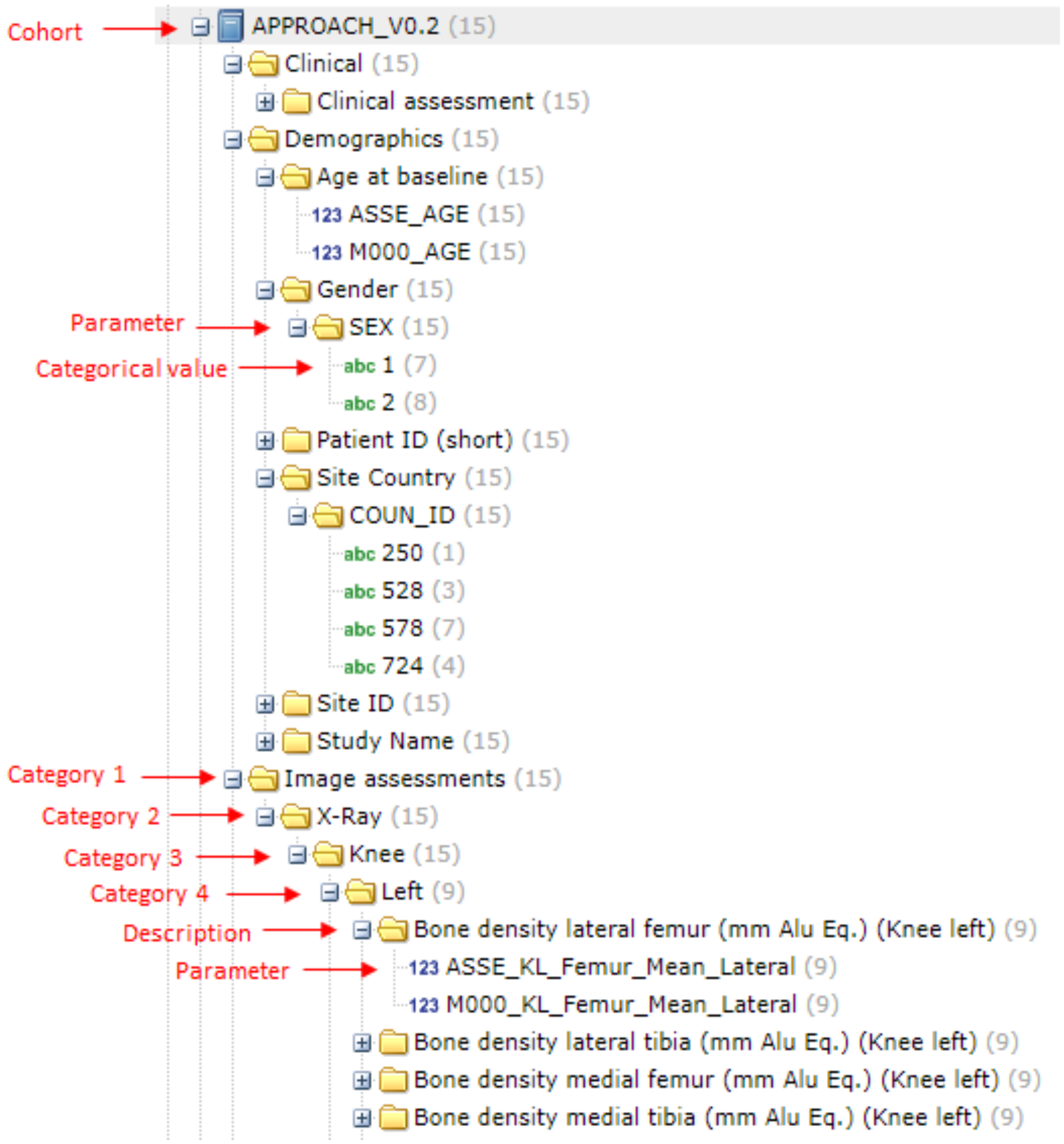


Figure 9: Screenshot of the APPROACH tranSMART hierarchical tree. Also indicating the codebook mapping.

The data files provided to the tranSMART-batch pipeline should be delimited text files (preferably tab delimited). Per file only data of one visit should be present and the parameters should be prefixed with the visit prefix (see last “Parameter” example in Figure 9). The first line of the data file should contain the parameter names (these should match with the parameter names in the tranSMART Codebook file). All subsequent lines contain the data.

The APPROACH pre-processing pipeline will convert the source data files into the format described above, using the tranSMART codebook to replace the source parameter names with the parameter names as listed in the tranSMART codebook. Besides converting the source data files in the correct format, the APPROACH pre-processing pipeline also performs QC on the data (based on information given in the codebook). All the pipeline scripts will create log files indicating what has been processed or QCed.

Also, in the APPROACH project data from retrospective cohorts were used to generate Machine Learning models. These models were used to rank the selection of patients for the prospective APPROACH cohort. The data of these retrospective cohorts were also uploaded into tranSMART. Before the upload of these data, a manually evaluation of the integration process for the parameters was performed. This evaluation highlighted that a wide range of metadata requires manual methods to be integrated. To solve this, a harmonization model composed by the following steps was suggested:

1) a separate mapping assessment for each parameter;
2) a cross-sectional evaluation of some parameters (within groups/class/index), where possible;
3) a selection of standardized tools for interoperability and harmonization process (i.e. ontology: HPO);
4) a comparison among potential harmonization solutions (in case of one-too-many mapping) capable to single out the proper tool for data capture and integration.

Despite that this approach is time consuming, this model provides a potential solution for harmonization of existing cohorts.

Conclusions

The main conclusion from the APPROACH use case is that it is crucial from the start to identify all data sources and the corresponding data flows into the tranSMART database. Ideally this should be described in a data management plan written at the start of the project. Based on this data flow overview the transfer steps will be worked out, including what tools and processing should be set up. Furthermore, the parameters captured should be described in a codebook or data dictionary. Working this way will make it possible to standardize the upload into a central database to facilitate multimodal biomarker research. Nonetheless, to a certain extent some custom work will always be required. The approach taken in the APPROACH project can serve as a blueprint for other projects. Even for those projects that will not be easily standardized the data will at least become more FAIR [19] thanks to the consistent inclusion of metadata.

Regarding retrospective cohort data, the APPROACH use case also shows us the need of combining an innovative approach with (limited) manual processing of data sets derived from retrospective cohorts. The identification of standards for the source data, in terms of vocabularies, ontologies and other standardization and quality tools, solves the difficulties arising during the integration process, guaranteeing fast and high resulting harmonization. This is the target for future research studies with a biomedical scenario, allowing the interoperability and, -when appropriate-, reusability of (cohort) data.

Cancer Core Europe use case

Another use case we set out to support was that of the international consortium 'Cancer Core Europe' (CCE), in which the following participants are involved: the Cancer Research UK Cambridge Centre ('CCC', UK), the German Cancer Research Center & National Center for Tumor Diseases (DKFZ/NCT Germany), Institut Gustave Roussy (IGR, France), Instituto Nazionale dei Tumori (INT, Italy), Karolinska Institutet (KI, Sweden), the Netherlands Cancer Institute (NKI, the Netherlands), and Vall d'Hebron Institute of Oncology (VHIO, Spain). These centres wish to collaborate on various translational and clinical research projects through effective data sharing and common database building. In this manner, a virtual e-cancer institute will be created, allowing the participating centres to carry out joint translational and clinical research, to conduct next-generation clinical trials, to develop personalised cancer medicine, to establish standardised academic diagnostic platforms, to create large shared databases, and to perform outcome research [3]. To achieve this collaborative data sharing there are seven task forces, focusing on different data types and aspects that are encountered in multi-centre studies..

NKI, on behalf of CORBEL, has taken the lead of the Data Sharing Task Force, and, together with the other participants, the 'Data sharing pilot BRCA1/2', was initiated in 2016 with the goal to demonstrate how effective data sharing may be achieved. There are three aspects to this data sharing pilot: 1) a common IT platform, 2) the actual use case content, and 3) the ethical/legal aspects.

A common IT platform

In 2016, a common IT platform was adopted within CCE, based on the applications available in the translational app store of the Dutch national research infrastructure Health-RI11 and CORBEL, allowing for enrichment, sharing and integration of different data types. During the project each participating centre would have to work out the technical aspects required for delivering data to central applications, from identification of eligible subjects to (automated) data publication on the selected IT platform. This project would serve to facilitate finding eligible subjects for future clinical studies and sharing of their data in future (CCE) projects, basically 'opening up the pipelines for future data sharing projects'.

In the pilot, clinical, biosample, and Next Generation Sequencing data are collected, shared and integrated, and four applications from the translational app store were selected for this purpose: OpenClinica, Molgenis Catalogue, tranSMART and cBioPortal.

- **OpenClinica** is an electronic data capturing system, and, through the use of an electronic Case Report Form (eCRF), clinical data from each of the participating CCE centres can be entered into this application using structured and coded forms. These clinical data encompass, amongst others, demographics, pathology data, epidemiological factors (e.g. environmental exposures, lifestyle factors), clinical presentation, patient history, disease evolution, therapy type and response, and outcomes.
- **Molgenis Catalogue** is an online catalogue in which the study metadata can be entered, i.e. general information and descriptions of a study or biobank and related access protocols to samples and data, and it is possible to capture detailed metadata of the biosample collection as well, providing insight into the availability of individual samples.
- **tranSMART**, as mentioned in the section 'Background' of this document, is a data-integration platform. For this use case, all 'final' or 'processed' clinical, biosample and the Next Generation Sequencing mutation status for BRCA1 and BRCA2 are to be entered into tranSMART. The views and questions that can be answered in tranSMART are, in particular, at the subject-group level, e.g.: for patients with a mutation in BRCA1 and/or BRCA2, what does their overall survival look like compared to patients who are wild-type for those genes?

[11] https://www.health-ri.nl/

Or: for patients who are older than 60, with a BRCA1 mutation and overall survival longer than 5 years, do we still have DNA biosamples in order to perform additional experiments?

- The **cBioPortal for Cancer Genomics**, as mentioned in the section 'Background' of this document, is another data-integration platform. For this use case, all 'final' or 'processed' data as described above are to be entered into cBioPortal. The views and questions that can be answered in cBioPortal are, for this use case, at the gene-level, e.g.: for patients with a mutation in BRCA1, what are the exact mutations reported? Or: are there any existing drugs (FDA approved or experimental) linked to this mutation?

Data loading, training

In the context of the collaboration between CORBEL and Cancer Core Europe, steps have been undertaken to standardise the procedures for data loading into tranSMART and cBioPortal, allowing a data manager or data steward, without (extensive) knowledge of programming languages, to import data into these two platforms. For OpenClinica and Molgenis Catalogue, the user manuals were already in place as well as the methods for data import suitable for easy upload.

For CCE, participating members have received several instructions and demos on the various applications and, in April 2019, they were invited to participate in a two-day workshop. During this workshop, hands-on experience with importing data to all four of the selected applications was obtained, and at the end of this two-day training the participants could import data into these applications. All training material was disseminated within the CCE Data Sharing Task Force and is available upon request.

Use case content of the 'Data sharing pilot BRCA1/2'

In this data sharing pilot, the focus is on collecting data from cases, pan-cancer, in which a patient has a somatic BRCA1 or BRCA2 mutation. Each centre has to obtain local permission, from either an Institutional Review Board or an Ethical Committee, to share eligible cases.

The initial study protocol stated as sole objective: 'to serve as a pilot for effective data sharing', by collecting different types of data and working out the method to share these in a harmonised and standardised fashion (one that could be automated, thereby facilitating future research studies as well). However, the study protocol had to be amended in Q4 2019 to include more clinically relevant objectives in order to obtain the required permissions, i.e. 'to investigate the prevalence of somatic BRCA1 or BRCA2 mutations'. For this current study protocol, two out of seven centres have obtained permission, while the other centres are awaiting approval (January 2020).

Pending approval, effort has been put into setting up the various codebooks and pipelines with which to collect the different types of data in a semantically harmonised manner.

For the collection of clinical and pathology data, a codebook and corresponding electronic Case Report Form in OpenClinica was created. Each CCE centre will locally work out how to supply the requested data in this manner, and import the data into OpenClinica. A similar approach will be taken for the biosample data, when more details become available.

For mutation data obtained through Next Generation Sequencing, the process to get from 'raw' sequencing data to harmonically annotated 'processed', or 'called' data is being worked out, and a method used in another CCE study is being adopted for this. This 'processed' sequencing data, combined with the other 'processed' data types, will be imported into the data-integration platform(s) tranSMART and/or cBioportal.

Ethical/legal aspects

One of the intentions of Cancer Core Europe is to create a virtual e-hospital and to establish a platform supporting common database building and data sharing, covering all aspects that are encountered in research, thereby facilitating the execution of multi-centre translational studies.

Ethical, legal and financial issues (ELFI) are always relevant when performing any type of research study, and are thus also relevant in CCE and any projects being performed between these participants. For this purpose a specific ELFI task force was created. Through the data sharing pilot BRCA1/2 such ethical and legal issues were going to be identified and solved, creating the basis for future studies performed within the CCE consortium.

The manner in which data is going to be shared is of concern here as well, and in order to use the common IT platform for data sharing (that is described in this report), each centre needs to conclude a Data Processing Agreement and a Service Level Agreement with Stichting TraIT, the legal entity hosting the chosen applications. For this purpose, the centres have examined the specific services of Stichting TraIT for tool security, conditions of use, and GDPR compliance. As of January 2020, four of the seven CCE participants have signed these particular agreements, meaning that pending local ethical approval, these centres can enter their data in these central, online platforms. However, integrating data from multiple centres or accessing another centre's data requires additional agreements relating to policies and terms of use on the to-be-shared data. These terms will be covered in a Data Transfer Agreement (DTA) and/or a Material Transfer Agreement (MTA). While there will always remain study-specific clauses, an 'umbrella' template could be created for these agreements, containing standard paragraphs to include with some optional additional text, depending on the type of study and data being collected and shared. This 'umbrella' document could speed up future studies and collaborations for which the same type of agreements needs to be established.

As CCE is not yet a legal entity entitled to sign agreements, this means that for each individual study and each individual centre, a DTA and/or MTA need to be drafted and signed, which is very time consuming. In order to simplify matters, a consortium agreement needs to be in place, mandating CCE to sign on behalf of its participants. Similar issues are seen in other consortiums as well, such as the IMI APPROACH project discussed above. As of writing this report (January 2020), the consortium agreement is in the 'finalising stages' of being signed.

Thus, the road forward for the data sharing pilot BRCA1/2 is:

- Each centre needs to gain local approval for sharing of patient case records, for the amended protocol, via either an Institutional Review Board, or an Ethical Committee.
- Each centre needs to sign a Data Processing Agreement and a Service Level Agreement with Stichting TraIT, the legal entity hosting the chosen applications, before the data can be entered into these applications.
- Each centre needs to upload their data as a separate study, working out the technical issues and the workflow to deliver the requested data, until such time as a Data Transfer Agreement between all the different centres has been arranged (which is dependent on the consortium agreement as well).

Conclusions

- The IT infrastructure for data loading into the selected central applications is in place, and CCE members have been supported in the context of this CORBEL use case to upload data into these applications, through demos and hands-on training.
- To facilitate research studies initiated within a consortium, it is important that the consortium has all the necessary policies and agreements in place, enabling a top-down approach. Once standardised document templates and procedures as well as IT infrastructure solutions are available, studies will be enabled to start sooner, allowing researchers to spend less time and effort on building customised solutions.
- Initiating a pilot use case to demonstrate how effective data sharing may be, has led to the identification of various pitfalls that each centre runs into. In particular the ethical/legal issues, but also aspects related to sharing of semantically harmonised data and their respective formats. For the latter, input and coordination is required of a domain expert, who is (made) aware of existing standards.
- Dedicated use-case assigned personnel to work out the various aspects of a use case or study is crucial.

Recommendations

Based on our experiences in the four use cases described above, we outlined a set of general recommendations below, followed by specific recommendations regarding the three key components of the infrastructure (tranSMART, XNAT, cBioPortal).

General

- Infrastructure for integrated analysis of multimodal data should be implemented in a modular way rather than with big monolithic solutions. Even within a single research domain, requirements, standards and best practices for data storage and analysis are evolving quickly. Integration of multimodal data brings together multiple domains, so this scenario requires even more flexibility. By aiming at an ecosystem of interoperable tools, selected components can be replaced, with relatively low effort, by different/newer/better solutions once they become available.
- Tools for data storage in medical research should implement an API to enable data access by automated processing software.
- Fine-grained control of access rights (who is allowed to read/edit/delete which data elements) is strongly advised, since the (personal) data are often highly privacy-sensitive.
- For use cases dealing with privacy-sensitive data, it should be possible for researchers to deploy the entire infrastructure within their own institute, hosted on their own storage & compute facilities within their firewall (or a private cloud). Preferably, this should be supported with a containerization technology allowing easy installation of those infrastructures.
- Domain experts should be involved in the data harmonization process from the start.
- Ethical and legal prerequisites require careful planning and an early start. In particular for large consortia, the legal entity entitled to sign the required documents is not always clear.

Planning the legal framework from the start and including the right clauses in e.g. a consortium agreement could avoid major legal obstacles at a later stage.

tranSMART

In the two use cases applying the tool (Cancer Core Europe and APPROACH), tranSMART proved to be a powerful tool integrating clinical data with multi-omics data, allowing for subcohort selections across the study. On the other hand, there were also some concerns about the usability and sustainability of the tool. This led to the following recommendations:

- Create a diagram visualizing the flow of data from source to central tranSMART database. This will be of great benefit for writing a data management plan.
- Set up a codebook/data dictionary describing all data collected with the accompanying agreements on the terminology to be used; apply ontologies wherever possible. Involve disease experts in this data harmonization effort; apparently identical parameters might be actually very dissimilar due to varying experimental conditions which is not easily picked up by non-experts.
- Data harmonization across the study will help to make the data more FAIR, and allows for automation of data migration pipelines offering technical interoperability. More generically, agree on the procedures for sending the source data files and make sure the agreed format of the data files remains unchanged. This ensures that the preprocessing pipeline will remain running without errors.
- Also use the codebook to configure the tranSMART hierarchical tree and to perform some basic QC on the source data, which will significantly improve the data quality and reliability.
- It is beneficial to allocate trained support resources for pre-processing the data. Preprocessing the data is not trivial requiring in-depth knowledge of the ETL tools in use, as well as basic programming skills. This can usually not be delegated to the researchers/clinicians in a study.
- In the project use cases we had to retro-fit the codebook based on existing data collections. In future projects it would be preferred to make use of standard disease-specific ontologies and vocabularies (if available) and use these ontologies to set up the codebook and database system. This will make it much easier to combine data from multiple cohorts and would require much less harmonization efforts.
- With regards to improvements to tranSMART itself, two main suggestions arose:
- Make tranSMART suitable for longitudinal data
- Make incremental data uploads possible
- It depends on the data types and the questions to be answered in a study which version of the data integration platform is suitable to address the project’s research questions. In general:
    - For subject or cohort level queries and analyses, tranSMART 16.2 is suitable as a platform. NB, this version is succeeded by more recent 18.X and 19.X versions, which were altered in such a way as to become more compatible with i2b2.
    - For sample and longitudinal queries of data, tranSMART 17.1 or i2b2 could serve as a data-warehouse. There are multiple 17.X versions, the most promising one being a version released with a new user interface called ‘Glowing Bear’. While this platform

    - has the capability to filter and query longitudinal data, it lacks some of the data integration capabilities with high-dimensional data of tranSMART 16.2.
    - For data integration and analysis of sample, longitudinal, and/or 'omics' types of data, another platform, such as cBioPortal, could be a suitable alternative.
- Within CCE (or any other multi-centre collaboration), each centre should be able to import data to the chosen applications. For this it is recommended that the data loaders participate in training sessions and share data uploading expertise between centres.
- Because there are multiple versions of tranSMART maintained in the community (see above for an overview), it is recommended to carefully observe the tranSMART community before investing any time, effort, and finances into additional functionalities. In the interim, cBioPortal could be considered as an alternative data-integration platform for many of the use cases.
- When setting up a data sharing platform to enable multi-centre collaboration, map out the process flow and agree on harmonised and/or standardised data formats. This will facilitate technical interoperability, and enable data to 'flow' from one application to another, should there be a wish to move the data from one platform to another.

XNAT

XNAT was used in three of the use cases and has proven itself as a reliable and powerful tool, both for the end user (clinical research-oriented scientists) as well as for more data science-oriented users (e.g. radiomics). During the execution of the use cases some specific learnings came up which have been summarized in the recommendations below:

- The XNAT API wrappings provided by Python libraries XNATpy[12] and PyXNAT[13] greatly facilitate programmatic interfacing with XNAT, as demonstrated both in the preclinical use case and the clinical PROOF use case.
- Based on our experiences in the APPROACH use case and several other multi-center imaging studies, a recurring challenge is the transfer of image data from the local institute's Picture Archiving and Communication System (PACS) to the central XNAT archive. The Radiological Society of North America (RSNA) recommends the use of the open source Clinical Trial Processor (CTP) software to facilitate this procedure. Although this software indeed greatly streamlines and standardizes the process of DICOM image anonymisation and data transfer, it requires installation by local IT personnel, who are not always available or willing to assist due to other priorities. We therefore recommend to contact local IT personnel and/or trial management offices as early as possible, and allocate sufficient resources in project budgets for this step.
- XNAT includes a built-in javascript-based image viewer, which implements basic viewing functionality, sufficient for first quality assurance, visual assessment, and data exploration. However, in many studies, more advanced and specialized visualization (and annotation) functionality is demanded. To satisfy this demand, XNAT can be connected to external viewers, either using the DICOM interface or the (HTTP-based RESTful) API, or simply by manually downloading the images from XNAT to a temporary disk location and opening from

[12] https://xnat.readthedocs.io

[13] https://pyxnat.github.io/pyxnat/

there. For new projects, we recommend to discuss the requirements and expectations with regard to image visualization at an early stage, allowing sufficient time to setup the desired pipeline.

- Although XNAT imposes a general structure (based on the hierarchy Project->Subject->Experiment->Scan), further harmonization of naming conventions is not imposed, and is the responsibility of the user. To enable linkage with other non-imaging data stored outside XNAT, the user should make sure to use corresponding subject IDs across platforms, or at least keep the key to convert between IDs. Harmonization of naming conventions within scan sessions is also recommended. For example, most MRI protocols contain different imaging sequences, whose names vary among MRI vendors and institutes where the data is acquired. XNAT offers the "Scan Type Cleanup" function, which simplifies relabeling scan types according to a systematic naming scheme. Finally, XNAT offers several possibilities to store derived images (for example conversion to NIFTI format or segmentations). Consequently, within each project, a consensus must be reached on where to store such images in the file hierarchy.
- The use cases considered in this project all assume that image data is stored in a single "central" XNAT instance. For multi-center studies, a federated approach is another possibility, where each institute involved in the study installs its own XNAT server and uses that to store the data. An advantage of such an architecture is that the data remains at the local institutes. However, this raises new challenges for analysis. It means that image analysis pipelines need to run locally as well, which is possible, but demands more support of local IT personnel, which may not always be available. Therefore, unless such support is available, we recommend to adopt a centralised architecture.
- At preclinical level several image raw proprietary formats exist, and DICOM converters are not provided by all the vendors or the stored DICOM information differ with no consistency across imaging modalities or inside the same imaging modality. Therefore, efforts are needed for the harmonization of the DICOM information stored inside the images.
- Recent modalities (i.e. Optical Imaging, Photoacoustic Imaging, Near-InfraRed Imaging) have not yet a dedicated DICOM tag to specifically identify these techniques. We recommend to include these new techniques in the DICOM standard for a proper identification of the medical images obtained with these scanners.

cBioportal

Although the execution of the Cancer Core Europe use case is still to be completed before the final results can be obtained, the usage of cBioPortal for cancer genomics use cases appear to be very promising. In the course of the implementation of the use case we derived some recommendations:

- Researchers are still largely unaware of the existence of cBioPortal, or that it may be suitable for studies with 'omics' data across disease areas, not just cancer-specific studies. Creating awareness of this tool in different institutes (newsletters, demos, training sessions) is therefore recommended.

- On the website of cBioPortal14 a lot of useful information is given in the FAQ, yet be sure to visit the tutorials tab on this website as well, as these give a quick insight into the portal's capabilities!
- cBioPortal has a good description on how to provide various data types15; with this, a bioinformatician is usually capable of supplying the correct data format for upload.
- For studies that have sample-level, 'omics' and/or longitudinal data aspects, cBioPortal can be the data-integration platform; if a bioinformatician wishes to do an analysis that is not present in this platform, the data of the selected cohort(s) can be exported for e.g. subsequent analysis in R.
- If a data manager or data steward wishes to learn how to import data into cBioPortal, general training material can be shared upon request. In addition, in the coming months we intend to organize an on-line training provided sufficient users show interest in this event.

Next steps

Since this report completes CORBEL task 3.4 there are not really any next steps within the context of the CORBEL project, except for the finalization of the support for some of the use cases. Within the Horizon 2020 project EOSC-life further recommendations and guidelines will be developed for the "cloudification" of the IT framework for multi-modal studies. The recommendations from this report will be the starting point for these EOSC-life activities.

Euro-BioImaging has been granted the legal status of an ERIC (European Research Infrastructure Consortium) on 29 Oct 2019. Within this framework, we will further maintain, support, and develop the solutions for imaging. Specifically, the developments on infrastructure for clinical and preclinical imaging will continue as part of the Euro-BioImaging Population Imaging Flagship Node (Rotterdam, NL) and the Euro-BioImaging Medical Imaging Hub (Torino, IT), respectively.

Delivery and schedule

This delivery is delayed, as approved in the frame of the 3rd Grant Agreement amendment, in line with the overall CORBEL project extension. This way, we could incorporate the latest results of the use cases in this report.

References

1. Runhaar J, van Middelkoop M, Reijman M, Willemsen S, Oei EH, Vroegindeweij D, van Osch G, Koes B, Bierma-Zeinstra SM (2015). Prevention of Knee Osteoarthritis in Overweight Females: The First Preventive Randomized Controlled Trial in Osteoarthritis. *Am J Med,* **128** (8), 888 - 895.
2. http://www.approachproject.eu/
3. Eggermont AM, Caldas C, Ringborg U, Medema R, Tabernero J, Wiestler O (2014). Cancer Core Europe: a consortium to address the cancer care-cancer research continuum challenge. *Eur. J. Cancer* **50** (16):2745-6. doi: 10.1016/j.ejca.2014.07.025.
4. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N (2012). The cBio cancer genomics

[14] https://www.cbioportal.org/

[15] https://github.com/cBioPortal/cbioportal/blob/master/docs/File-Formats.md

portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov;* **2** (5); 401–4

5. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* 6 (269):pl1
6. Muzny DM *et al,* Comprehensive molecular characterization of human colon and rectal cancer (2012). *Nature* **487**; 330-337. PMID: 22810696
7. Marcus, D.S., Olsen T., Ramaratnam M., and Buckner, R.L. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data (2007). *Neuroinformatics* **5**(1): 11-34
8. D. Mason, SU-E-T-33: Pydicom: An Open Source DICOM Library (2011) *Medical Physics* **38** (6): 3493
9. Schwartz Y, Barbot A, Thyreau B, Frouin V, Varoquaux G, Siram A, Marcus DS, Poline JB. PyXNAT: XNAT in Python (2012), *Front. Neuroinform.* **6**: 12.
10. **IEEE 2019** International Symposium on Biomedical Imaging (ISBI'19), Venice (Italy), 8-11 April, 2019 – "A Customizable Workflow Engine to Store, Process and Share Medical Images for Preclinical Imaging Centers" and **EMIM 2019** 14th European Molecular Imaging Meeting, Glasgow (United Kingdom) 19-22 March 2019 – "Developing a customizable workflow engine for storing, sharing, processing and reusing medical images for preclinical imaging facilities"
11. Delude CM, Deep phenotyping: The details of disease 2015). *Nature* **527**:S14-5; https://doi.org/10.1038/527S14a.
12. Lambin *et al* (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* **14** (12):749-762. doi: 10.1038/nrclinonc.2017.141.
13. Xu D, Hansson M, Klein S, Oei EH, Wagner F, Bierma-Zeinstra SM, Runhaar J, Association between meniscus volume and development of knee osteoarthritis. *Osteoarthritis and Cartilage* 27, S272-S273. https://doi.org/10.1016/j.joca.2019.02.650
14. Hirvasniemi J, Klein S, Schiphof D, Oei E, Discrimination of subjects with and without osteophytes using a radiomics approach on tibial bone, ESMRMB 2019.
15. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, Klaver CCW, Nijsten TEC, Peeters RP, Stricker BH, Tiemeier H, Uitterlinden AG, Vernooij MW, Hofman A. The Rotterdam Study: 2018 update on objectives, design and main results (2017). *Eur. J. Epidemiol.* **32**(9):807-850
16. tranSMART batch documentation: https://github.com/thehyve/transmart-batch/tree/master/docs
17. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis (1957). *Ann Rheum Dis.* **16**:494–502. doi: 10.1136/ard.16.4.49
18. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2) (2010). *J Am Med Inform Assoc.* **17**(2):124-30.
19. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship (2016). Sci Data. **15** (3):160018. doi: 10.1038/sdata.2016.18.

Abbreviations

API	Application Programming Interface
APPROACH	Applied Public-Private Research enabling Osteoarthritis Clinical Headway
CT	Computed Tomography
CEST	Chemical Exchange Saturation Transfer
CORBEL	Coordinated Research Infrastructures Building Enduring Life-science services
DICOM	Digital Imaging and Communications in Medicine
DWI	Diffusion Weighted Imaging
ESFRI	European Strategy Forum on Research Infrastructures
eTRIKS	European Translational Information & Knowledge Management Services
ETL	Extract - Transform - Load
FAIR	Findable, Accessible, Interoperable and Reusable
GEO	Gene Expression Omnibus
HGNC	HUGO Gene Nomenclature Committee
HPO	Human Phenotype Ontology
IMI	Innovative Medicine Initiative
JSON	JavaScript Object Notation
MAF	Mutation Annotation Format
MIG	Minimum Information Guidelines
MRI	Magnetic Resonance Imaging
NGS	Next Generation Sequencing
OA	Osteoarthritis
PET	Positron Emission Tomography
PROOF	Prevention of Knee Osteoarthritis in Overweight Females
QIB	Quantitative Imaging biomarkers
REST	Representational state transfer
RI	Research Infrastructure
SNP	Single-nucleotide polymorphism
TraIT	Translational Research IT
US	Ultrasound
VCF	Variant Call Format
XML	eXtensible Markup Language
XNAT	Extensible Neuroimaging Archive Toolkit