European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.12 – Final report on Data Curation

Due date of deliverable: 30th September 2017

Actual submission data: 14th September 2017

Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including Commission Services)	
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

DELIVERABLE INFORMATION

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	115446
Document	
Deliverable number:	D4.12
Deliverable title:	Final report on Data Curation
Deliverable version:	1.0
Due date of deliverable:	30th September 2017
Actual submission date:	14th September 2017
Leader:	Reinhard Schneider, Manfred Hendlich
Editors:	Reinhard Schneider
Authors:	Adriano Barbosa
Reviewers:	Jay Bergeron, Chris Marshall, David Henderson, Francisco Capdevilla Bonachela
Participating beneficiaries:	UL, ICL
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Manfred Hendlich and Reinhard Schneider
Work Package participants:	Adriano Barbosa; Wei Gu; Venkata Satagopam; Emmanuel Van der Stuyft; Francisco Bonachela-Capdevila; Bertrand De Meulder; Kavita Rege
Estimated person-months for deliverable:	20
Nature:	Report
Version:	1.0
Draft/Final:	Final
No of pages (including cover):	25
Keywords:	Curation, Data

Final Report on Data Curation

1 Abstract

The purpose of this document is to provide a brief compilation on the curation efforts lead by eTRIKS WP4 over the course of the project. Therefore, the report will compile information disseminated in the 1st, 2nd 3rd and 4th progress reports (deliverables D4.5, D4.7, D4.9 and D4.11, respectively).

2 Public server

Contributors: Kavita Rege (UL), Wei Gu (UL), Adriano Barbosa (UL), Venkata Satagopam (UL)

2.1 Introduction

The prime objective of the eTRIKS public server, as described in deliverables D4.5 (1st Progress report on Data Curation, section 1.2 "Aim of the Public server delivery package"), is to give access to curated and standardized public studies through public eTRIKS/tranSMART server. The main task of the eTRIKS public server is to consider, curate and make publically accessible those public studies that are of interest to different IMI projects. We also apply eTRIKS data curation and quality standards to these public studies so as to facilitate the integrated analysis of these studies in the eTRIKS tranSMART software.

In the following sections, the 1.2 Description section deals with Gene Expression Omnibus (GEO)[1] database from where the public studies are retrieved. Section 1.3 deals with the types of data that are fetched and curated from the GEO database. The 1.4 section provides the detailed methods, Algorithms and pipelines used. In 1.5 we discuss problems encountered during data curation and upload.

2.2 Description

According to [Barrett et al., 2013] *"The Gene Expression Omnibus(GEO)[2] is an international public repository for high-throughput microarray and next-generation sequence functional genomic data sets submitted by the research community."*

This database consists of more than 32000 public series comprising 800000 samples derived from more that 1600 organisms submitted by 13000 laboratories [Barrett et al., 2013]. The data is submitted to GEO in the form of three objects namely, Platform, Series, Samples. The database is extensively indexed; hence searching for relevant data becomes easy.

GEO database supports the bulk download of required gene expression studied from the GEO FTP site. This makes it easy to fetch data resource for tranSMART Dataset Explorer.

[1] http://www.ncbi.nlm.nih.gov/geo/
[2] http://www.ncbi.nlm.nih.gov/geo/

The tranSMART public server contains data collated from various sources such as NCBI GEO (Gene Expression Omnibus) (Figure 1), TCGA (The Cancer Genome Atlas) and the Gene Expression Atlas (ATLAS)

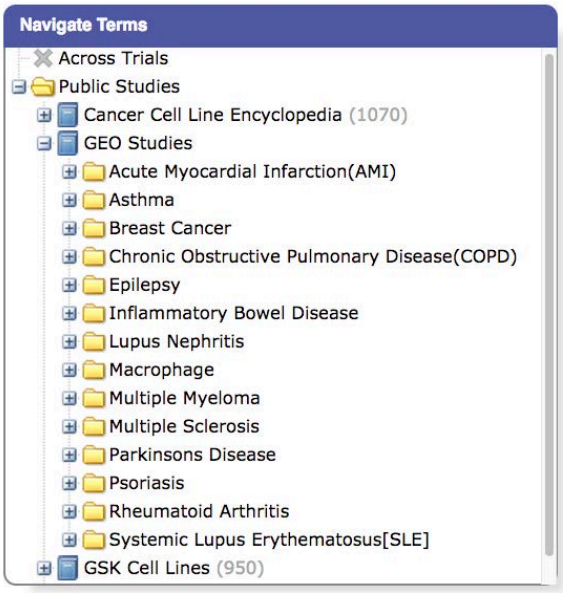


Figure 1: NCBI GEO Studies represented in the Public Server. 14 Diseases and a total of 10,687 samples are available (See Table 1).

Table 1: Distribution of studies and samples per disease in the Public Server.

Disease	#Studies (Samples)	Disease	#Studies (Samples)
Acute Myocardial Infarction	12 (998)	Macrophage*	19 (1,617)
Astma	25 (1,838)	Multiple Myeloma	1 (264)
Breast Cancer	3 (873)	Multiple Sclerosis	1 (94)
COPD	1 (58)	Parkinson's Disease	16 (496)
Epilepsy	1 (48)	Psoriasis	24 (989)
IBD	15 (849)	Rheumatoid Arthritis	23 (733)
Lupus Nephritis	2 (104)	SLE	27 (1,726)

Figure 2: GSK Cell lines source. 28 tissues/organs have been represented in the 950 cell line types deposited on Public Server. Number in parenthesis represents the number of cell lines per tissue.

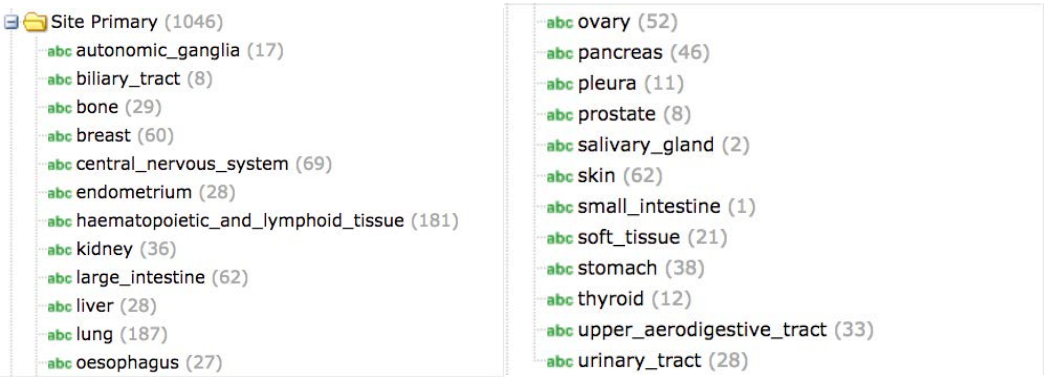


Figure 3: Cancer Cell Line Encyclopedia primary sites of 1070 sell lines. Number in parenthesis represents number of cell lines per primary site.

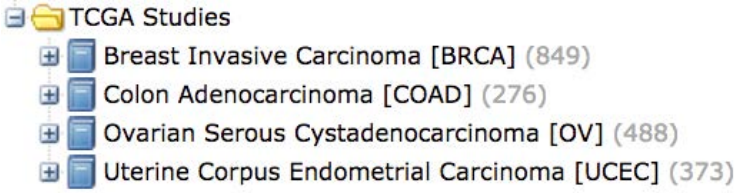


Figure 4: TCGA Studies represented on the Public Server. Number in parenthesis represents number of samples per study.

2.2.1 Data Acquisition Pipelines

Two pipelines for data acquisitions were developed as reported in the 1st Report (D4.5). These pipelines were used for extraction of data from NCBI GEO (Figure 5) and TCGA (Figure 6).

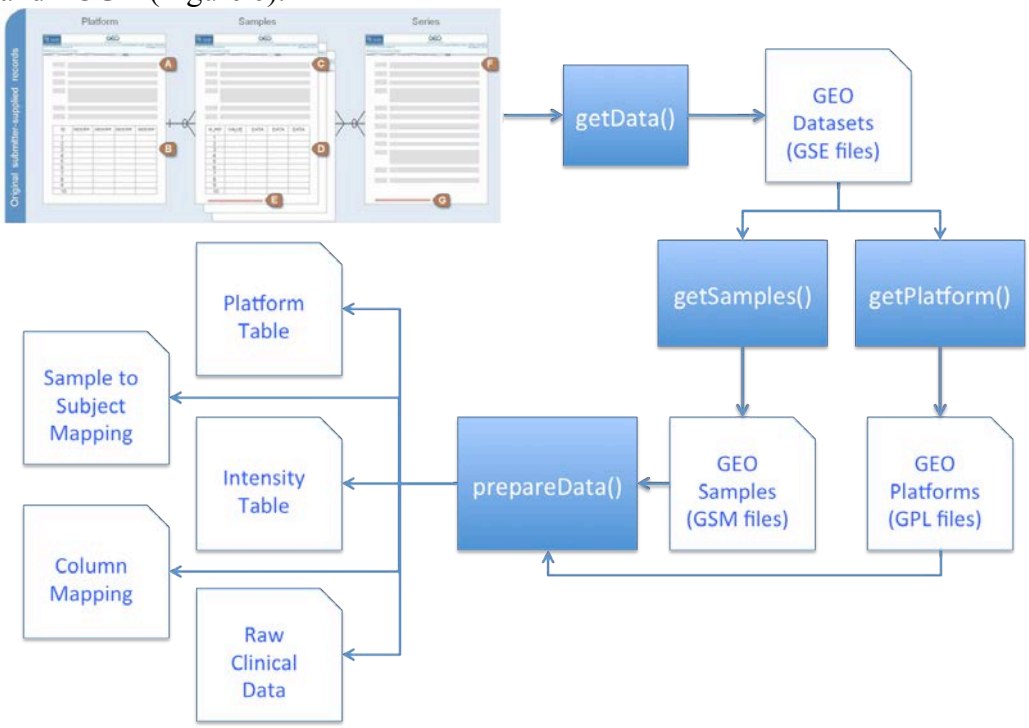


Figure 5: Simplified schema of GEO-Dataset Explorer pipeline. See description on deliverable D4.5

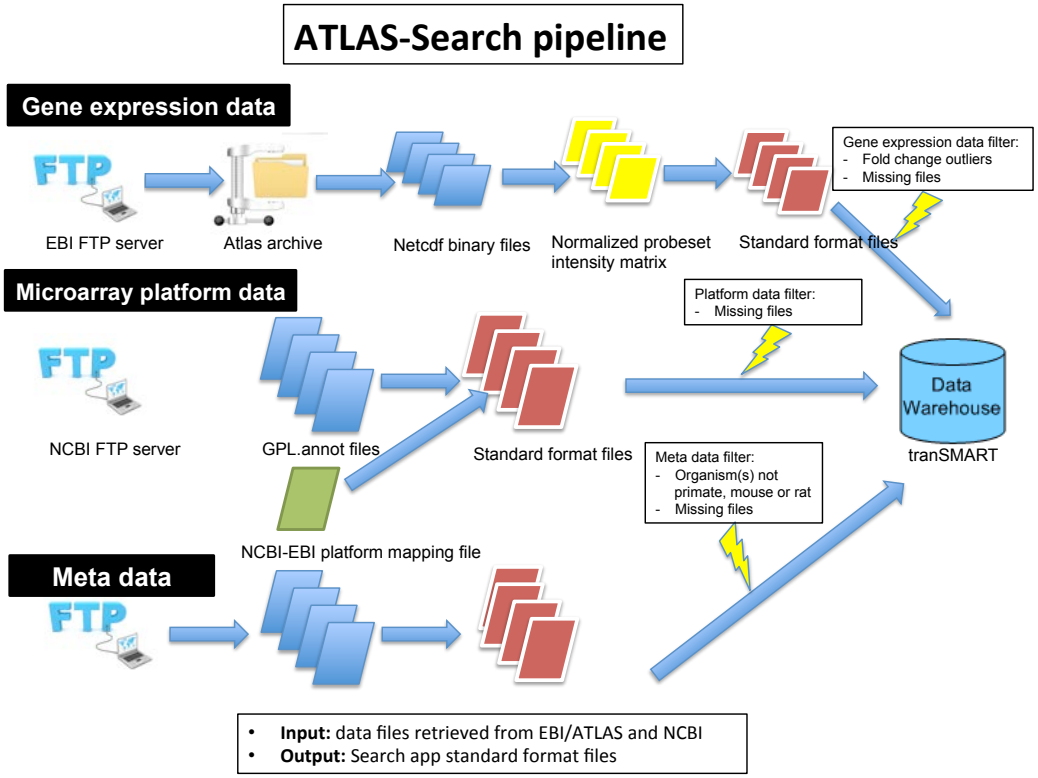


Figure 6: Simplified schema of Atlas-Search pipeline. See description on deliverable D4.5.

2.3 Problems Encountered

The major hurdle in uploading studies to tranSMART from the GEO database is that there is no standard followed by the data owners while uploading metadata. For larger numbers of studies, the metadata provided is inconsistent or incomplete for many key fields. Sometime these data are found in other fields. Hence, manual intervention of data processing for these fields makes the data curation task a tedious one. For example, in upgrading the public tranSMART server from version 1.2.2 to version 16.1, manual upload of all the previously loaded dataset was required.

3 Curation training and documentation

There was no new training since the last update report.

4 Overview on eTRIKS supported projects

This session compiles a list of the projects supported by eTRIKS. It includes the project description, curated data types, strategy for curation and problems encountered. This session might be overlapping with the previous reports, or referring to resources already described earlier.

4.1 APPROACH

Wei Gu (UL) and Andreas Tielmann (Merck)

4.1.1 Project description

IMI APPROACH aims to implement a comprehensive and high quality biomarker assessment to characterise osteoarthritis (OA) patient subsets and support future regulatory qualification and endpoint validation.

The project will provide a framework to identify the "right patient" to treat for a given drug by linking OA patient subsets to potential DMOAD targets based on phenotypic biomarkers, highlight specific disease drivers and progression criteria.

Finally, APPROACH wants to build a stronger collaboration within and among academic and industrial groups to enable future OA therapeutic development.

4.1.2 What data types have been curated?

Until the time of this deliverable, eTRIKS has been working on the curation of two publically available datasets:

- FNIH Osteoarthritis Biomarkers Consortium Project
- Cohort Hip and Cohort Knee (CHECK cohort)

For the FNIH cohort, there are 600 subjects, each with more than 350 variables collected. The first round of curation is finished with a full-dataset and a reduced-dataset (a subset filtered based on the full-dataset) both loaded to the APPROACH-tranSMART working server hosted at the University of Luxembourg.

For the CHECK cohort, there are 631 subjects. So far we have finished the curation of a subset of 16 variables. This subset has been also loaded to the APPROACH-tranSMART working server hosted at the University of Luxembourg.

4.2 AETIONOMY

Contributors: Adriano Barbosa (UL) and Wei Gu (UL)

4.2.1 Project description

AETIONOMY[3] is novel in terms of both, its scientific approach and its scale. There is a lot of published literature on the potential causes of Alzheimer's and Parkinson's disease and a significant number of major collaborations already funded and working well. The majority of these are looking at individual hypotheses or approaches to the problem e.g. genetic association studies, imaging studies, non-motor Parkinson's disease or familial Alzheimer's disease. Rather than start another similar approach, AETIONOMY will identify all of the available datasets either from published literature, publically available datasets or datasets from our collaborators.

A common framework will be developed which will allow the integration of data relevant for modeling and mining. Once this data has been curated (re-annotated and quality controlled) and put into the common framework, novel data mining and visualization approaches will be used to identify the pathophysiological changes occurring in the disease process at a molecular level. The knowledge extracted from the datasets will be used to cluster individual patients into separate mechanism based sub-groups leading to a new taxonomy of Alzheimer's disease and Parkinson's disease.

[3] http://www.aetionomy.eu

eTRIKS will mainly support the activities of AETIONOMY's WP2 to acquire, to curate and to build the data cube infrastructure which will integrate the available data. So far, the eTRIKS curation workflow has been adapted to cope with the curation needs of AETIONOMY.

4.2.2 What data types have been curated?

AETIONOMY will use tranSMART as one of the main components of the AETIONOMY Knowledge Base[4] (AKB). For that purpose, eTRIKS support is needed so set-up the AETIONOMY tranSMART server[5] as well as to load the selected studies to the system.

AETIONOMY has selected public studies that are relevant for its purposes and used some studies previously loaded at the eTRIKS public server have been re-uploaded to the AETIONOMY tranSMART server. Additionally, we provided support on the curation of a total of 13 datasets, a total of 7,337 samples (Figure 7)

Also 13 datasets (7,337 samples) have been curated as support activities for AETIONOMY (Figure 7).

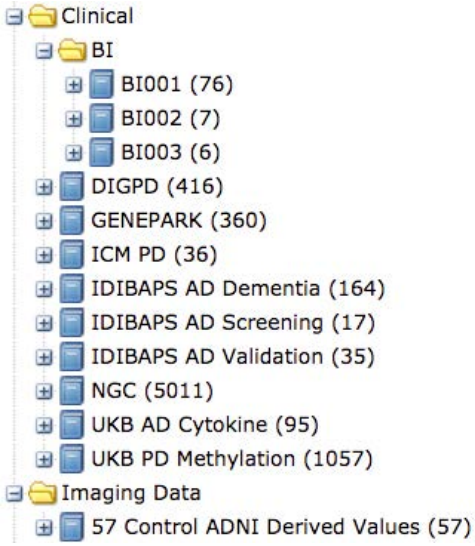


Figure 7: Curated datasets for the AETIONOMY project. These datasets are maintained in a dedicated AETIONOMY server.

4.2.3 What methods/algorithms and/or pipelines have been developed/used?

The methods applied to AETIONOMY are the same described for projects mentioned on previous reports.

4.2.4 What problems have been encountered?

No major problems once the PD Studies from eTRIKS were loaded to the AETIONOMY server. We also had to contact the data owners repeatedly in order to receive a precise description of the datasets to be curated.

[4] http://aetionomy.scai.fhg.de/
[5] https://aetionomy.uni.lu/transmart

4.3 RA-MAP

Contributors: Denny Verbeeck (JnJ) and Francisco Bonachela-Capdevila (JnJ)

4.3.1 Project description

RA-MAP[6] is a public-private collaborative project into early Rheumatoid Arthritis (RA).

RA-MAP seeks[4]:

- To identify the key predictors of clinical response and remission in RA patients, and;
- To identify those individuals at high risk of developing RA.

By understanding the human immune system in RA through the study of biological samples from RA patients we plan to develop an 'immunological toolkit' measuring the immune status of healthy individuals and patients.

The goal of RA-MAP[4] is to identify predictors of remission in RA.

There is a major need to identify the characteristics of those individuals most likely to achieve clinical remission so that both new and existing therapies can be targeted to the right patient populations.

4.3.2 What data types have been curated?

For the moment, mostly clinical data and gene expression data. In the future, small RNA and metabolomics data will be included.

Clinical data comes from the TACERA study and includes 273 patients. Since the project is still ongoing, the study in tranSMART is updated every three months. Currently, gene expression data available in tranSMART comes from the "Pilot" experiment, which includes a subset of 12 TACERA patients and 8 controls. More microarray data for TACERA patients are expected by the end of 2015.

The RA-MAP curated public studies have been loaded into the RA-MAP tranSMART instance. Studies with less than 10 patients have been filtered out.

4.3.3 What methods/algorithms and/or pipelines have been developed/used?

- The clinical data curation is at an early stage. Checks are performed to ensure that the provided data fall within valid types and valid ranges. It is also checked possible inconsistencies that might lead to data reformatting when necessary. Any reformatting is agreed with the data provider at King's College.
- Derived data columns are obtained based on the feedback of tranSMART users to increase cohort selection flexibility.

[6] https://research.ncl.ac.uk/mrgnewcastle/translationalprojects/ramap/

- As for gene expression, both raw and normalized data are uploaded to tranSMART. Raw data are extracted with GenomeStudio from the idat files provided by Tepnel Pharma Services and normalized data are obtained using neqc method in Limma/Bioconductor over the raw data.
- Currently, an R script is being developed to check that data uploaded to tranSMART are consistent with and equivalent to the original gene expression data. In this way, the data owner or data user can be sure that the data has not been altered while being uploaded.

4.3.4 What problems have been encountered?

Data curation is a slow process since it involves several actors. It needs to be ensured that any data changes produced during the curation process are agreed and approved by the data donor. It also needs to be ensured that any curation script is shared within the RA-MAP community.

4.4 ABIRISK

Contributors: Wei Gu, Nathalie Jullian, Fabien Richard and Serge Eifes

4.4.1 Project description

ABIRISK will have access to large cohorts of patients treated with biopharmaceutical products (BPs). Analyses of the mechanisms and consequences of immunization against biopharmaceutical products (BPs) require extensive post-marketing follow-up of patients, with comparisons of several BPs and various clinical conditions treated with the same BP. Sufficient numbers of patients must be included in each subgroup for the reliable evaluation of independent parameters. There is also a need for high-quality data generated by centres familiar with clinical research. The ABIRISK consortium has been designed to meet all of these requirements in order to target three types of disorders:

- Hemophilia A
- Multiple sclerosis
- Inflammatory diseases: inflammatory rheumatisms (including rheumatoid arthritis) and inflammatory bowel diseases

The ABIRISK Project collects data both retrospectively from patients suffering from various types of diseases and treated with various BPs at European centres with a high level of experience in clinical research and will prospectively recruit additional patients in dedicated studies during the 5 years of this program. Guidelines and Standard Operating Protocols (SOPs) for the study of anti-drug (AD) immunization will be established and used to standardize the collection of prospective data from these patients.

4.4.2 What data types have been curated?

Clinical data with multiple visits that has been provided by the Karolinska Institute (KI), Sweden and University of Innsbruck (UI), Austria have been curated so far. A more detailed overview on this data is given in **Table 2**.

Data provider	Wave	Number of variables	Number of patients	Number of visit records
UI	1	34	4500	12000
UI	2	93	4500	12000
KI	1	42	6300	16000

Table 2: Overview on clinical data curated for the ABIRISK project

4.4.3 What methods/algorithms and/or pipelines have been developed/used?

The University of Luxembourg (UL) team has developed a data quality-checking pipeline to check data integrity defined in the data control definition file. In the data control definition file, features of each column in the data file have been defined (e.g. data type, whether the column is compulsory or not). Then the data quality-checking pipeline will go through all the data columns in the data file and check if the data fulfil the definitions.

4.4.4 What problems have been encountered?

In ABIRISK tranSMART 1.1 server (hosted by eTRIKS), data export function could not generate correct output for visit of categorical data. Basically the visit information of categorical data is lost during the exporting. This has been found to be a general problem of tranSMART 1.1 version. A fix was provided by UL team, in detail a new exporting function is developed to add the missing visit information. This solution overcomes this issue in TM 1.1. The new TM version (1.2) has solved this problem.

4.5 OncoTrack

Contributors: Adriano Barbosa; Serge Eifes; Wei Gu; David Henderson; Gino Marchetti; Nathalie Jullian; Ioannis Pandis; Anthony Rowe; Venkata Satagopam; Emmanuel Van der Stuyft

4.5.1 Project description

The goal of OncoTrack is to identify and characterize biological markers that will help the understanding of the variable make-up of tumours and how this affects the way colorectal cancer patients respond to treatment. This question is being tackled with cutting edge laboratory-based genome sequencing techniques coupled to novel computer modelling approaches to study both the biological heterogeneity of colon cancers (i.e. patient to patient variability) as well as tumour variation within the

patient – for example, by comparing primary tumours with metastases. For more details see section "Objectives" on the OncoTrack website[7].

The results of this research is expected to allow the identification and qualification of a set of biomarkers that may be used to guide patient therapy, provide immediate feedback upon the effects of treatment and ultimately indicate likely outcome of disease management – i.e. Oncology Tracking!

4.5.2 OncoTrack data landscape

The OncoTrack data landscape is schematically represented in **Figure 1**.

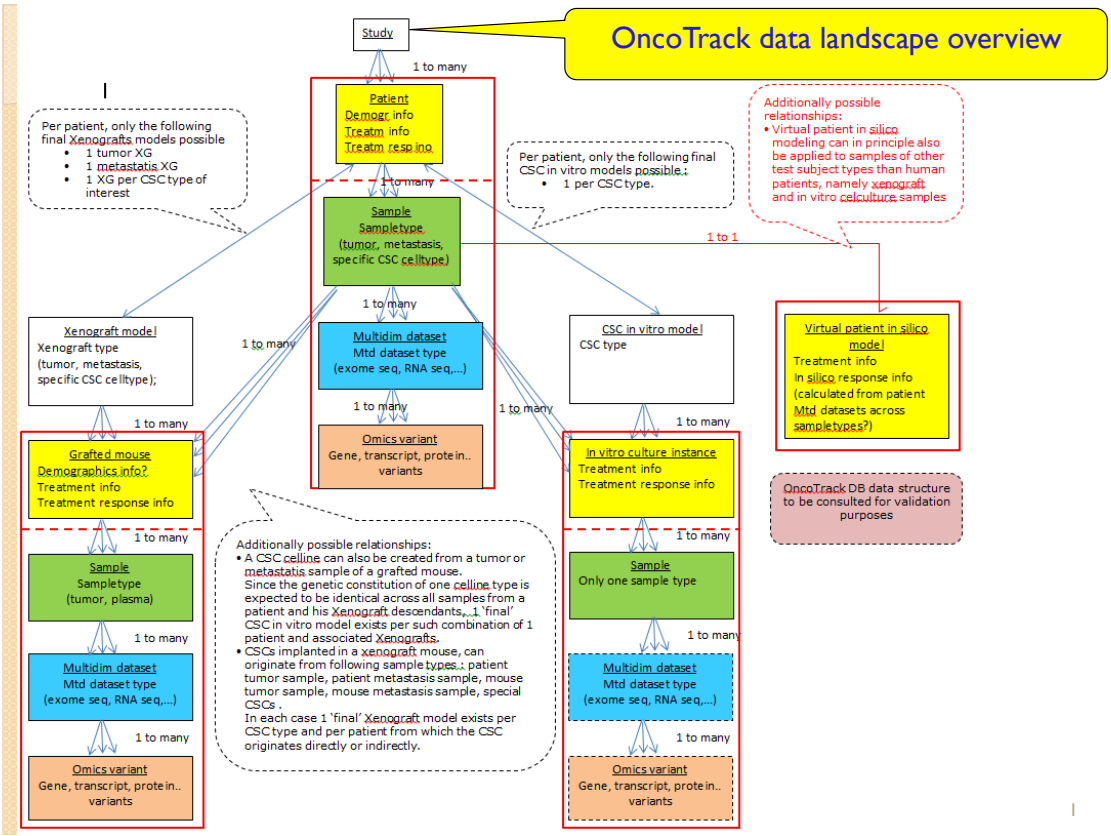


Figure 1: Schematic representation of the OncoTrack data landscape

4.5.3 What data types have been curated?

The low dimensional data currently available in OncoTrack for the different subject types involved in the research (patients, xenografts, cell-lines and in silico models) have been curated into separate studies for each subject type, together with reference data to capture the interrelationships between the different subjects (e.g. what is the parent patient from which a xenograft was created?)

So far, > 100 variables of > 300 Patients, > 400 Cell line treatment groups, > 750 Xenograft treatment groups and > 3000 *In silico* model treatment groups have been curated.

[7] http://www.oncotrack.eu

Given the importance of genomic variation in OncoTrack's research, work is currently ongoing to enable scientists to view - for a cohort selection of samples in tranSMART 1.2 - genomic variation information that can be captured in the VCF format. This is based on an integration with the Dalliance genome browser. A proof of concept is now available based on public VCF data. Integration of OncoTrack data is planned shortly.

Once the above functionality is available, we will reduce to a sub-selection only the overall mutation summary we experimented with in our earlier implementation in the low dimensional concept tree space. This allowed cohort selection based on whether any combination of genes of interest was mutated or not, but which in a full-sized experiment would likely over-stress the system by the sheer amount of concepts needed to cover that information.

Other multidimensional data being worked on includes methylation and RNA seq data.

- M-scaled differential Methylation values between an individual patient's tumor vs healthy sample on a probe-by-probe basis have already been successfully uploaded into tranSMART.
- Based on interaction with Lee Butcher (OncoTrack methylation expert) we have agreed to rather focus on Beta-scaled methylation values at a higher aggregation level (gene regions as well as 'differentially methylated regions' which may span across genes). We are awaiting sample data from Lee Butcher (University College London to experiment with their curation and upload.

- The curation of RNA-seq data into the OncoTrack tranSMART work environment has been successfully tested based on public RNA-seq data. As soon as such data are available from OncoTrack DB, these will be made available in tranSMART.

Patient clinical data are not yet available for reasons related to data privacy. As these become available soon, these too will be curated into tranSMART.

For each of the data types currently curated, an export script has been written to draw the relevant data from the OncoTrack data repository environment (OncoTrack DB) and to automatically transform them into the column-mapping format required for data upload into tranSMART. Since OncoTrack data are collected on an ongoing basis, a next logical step could consist in the automation of the data curation and upload process at regular time intervals.

4.5.4 What methods/algorithms and pipelines have been developed/used?

By adopting tranSMART 1.2 we get transparent access to:

- flexible data export functionality based on specifically selected cohorts/subsets. This cohort-based data can then be directly accessed by analytical tools outside of the tranSMART environment

- data import functionality into the Galaxy workflow environment which supports flexible data processing/analysis pipelines based on the exported cohort data from tranSMART.

Since it became very clear that the degree to which tranSMART will eventually add value to OncoTrack critically depends on the flexibility with which the interrelationships between different subject types in the overall OncoTrack data landscape can be integrated into querying and analysis. We tried to address the current tranSMART limitations in bridging across subject types in two complementary ways:

- Via the tranSMART API, and leveraging the above-referred referential data that provide the links between subject types.
  To illustrate the API capabilities, an R script is being developed to create a report of how different subject types align as to their response to similar treatment.
  This should illustrate the great potential in an interaction platform, where scientists can iteratively formulate their investigational questions and have these addressed by bio-informaticians, with API-based access to the tranSMART data warehouse.

- Via the development of a script to allow for the cascading from within the tranSMART web interface of different query filters across the different subject-types in the overall OncoTrack data landscape. Such scripting enables one to work through the following example query cascade:
  - In the xenograft study, create two cohorts based on treatment response:
    - Cohort 1: responders; Cohort 2: non-responders
  - In the corresponding patient study, create two cohorts to investigate patient-based genomic variations which could explain the response vs non-response in their corresponding child-xenografts.
    - Cohort 1: all parent patients of xenograft in the responder cohort
    - Cohort 2: all parent patients of xenografts in the non-responder cohort

4.5.5 What problems have been encountered?

The section here below gives an overview of the problems we have encountered with tranSMART version 1.1.

A core limitation we struggled with (as explained above) boils down to this:

- tranSMART can only capture 1 value per low dimensional tree concept. This was the cause why we had to cut up the study data into separate studies per subject type, since concepts for patient-derived subject types typically have multiple values for the same patient.
- Given its dependence on the i2b2 layer, the extension of tranSMART query filtering capabilities to include sub-lists of categorical variables seemed like a big problem. This workaround to the above problem was hence implemented in a script as explained above.

  Given the importance to OncoTrack of the above two bullet points, a set of slides was included in appendix to articulate the problem for which the team eventually found a workaround.

Also performance and stability issues were met quite often, both on the upload and tranSMART usage side.

Given the fact that API and VCF-genomic data support are crucial to OncoTrack, and missing in tranSMART 1.1, we decided to take the risk of becoming early adopters of tranSMART 1.2, with this and much other desirable extra functionality available.

On the critical path of making the above-described functionality operationally accessible to the OncoTrack scientists is the deployment of the Luxembourg-hosted test environment on the operational eTRIKS tranSMART hosting environment in Lyon.
Prerequisites for that are:

- Getting CDAs/MTAs in place between OncoTrack and CC-Lyon/eTRIKS
- The deployment of an operational eTRIKS tranSMART 1.2 version.

Quite some nice upfront work that could be done, independently of where these prerequisites are at, has been done by the Lyon team (including deployment of test environment and LDAP setup).

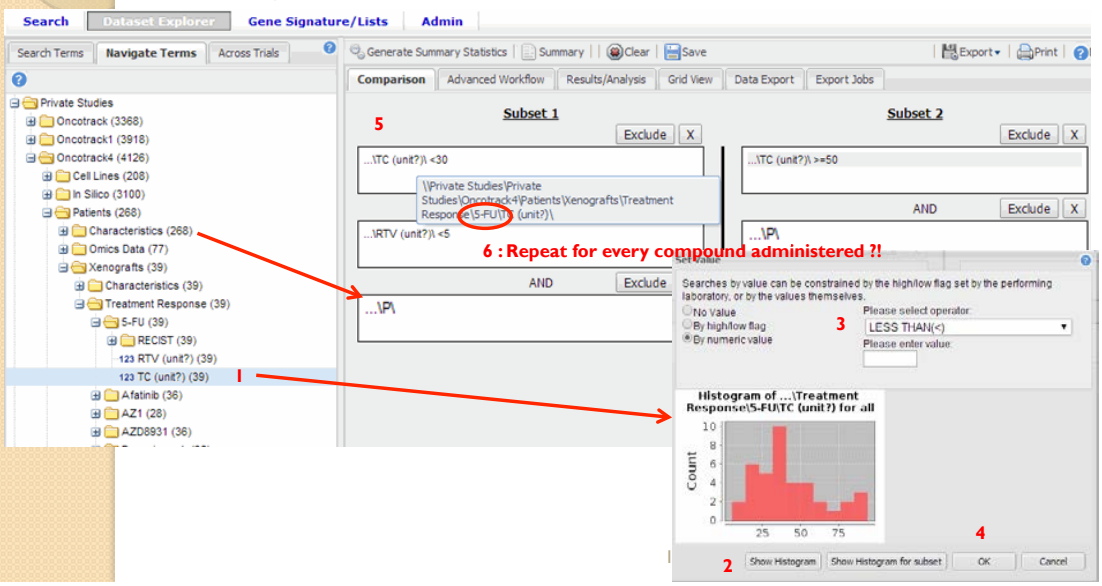
The data curation and upload work, which has been painful and slow in its initial stages, has picked up significant speed and effectiveness lately. Two key factors in that have been:

- The dedication of significant bandwidth by the Luxembourg team with great data curation and bio-informatics expertise in the OncoTrack data curation effort, as agreed to in the latest eTRIKS resource team meeting.
- The availability of tranSMART 1.2 functionality

4.5.6 Appendix: Articulation of a key-issue (addressed via scripting by Wei Gu, University of Luxembourg)

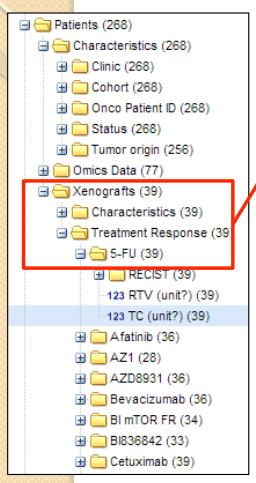
Flexible querying to hand over to analysis tooling

- **Question:**
  - Find distribution of Xenograft responders vs non-responders for patients with primary tumors
  - Assume responders -> ~~TC<30 AND RTV<5~~ RTV<=1.3 AND TC<50
  - Assume non-responders ~~TC>=50~~ TC>80



Data upload alternatives

Limitation : Per patient only 1 value per distinct variable



Multiple Xenograft model values per patient	**Only 1 retained per patient**
Multiple compounds administered per patient	**From generic 'compound administered' variable (= easy to query since representing a full collection of possible values)** **to 14 'hardcoded' sub-variables (= each variable contains only 1 value of possible compounds administered)**

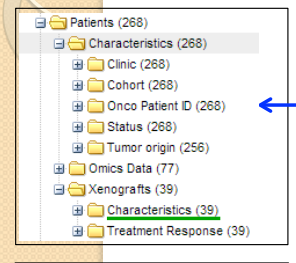
Impact
- **Generic querying compromised**
- **Effectively working with variables later compromised**

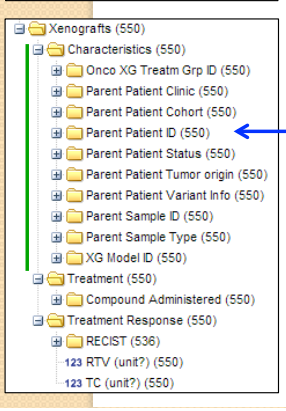
Variables	Cmpd admin	TC
Values	5-FU	20
	AZ1	15
	XYZ	8
	...	...

Variables	5-FU\TC	AZ1\TC	XYZ\TC	...
Values	20	15	8	...

Data upload alternatives

Limitation : Per patient only 1 value per distinct variable





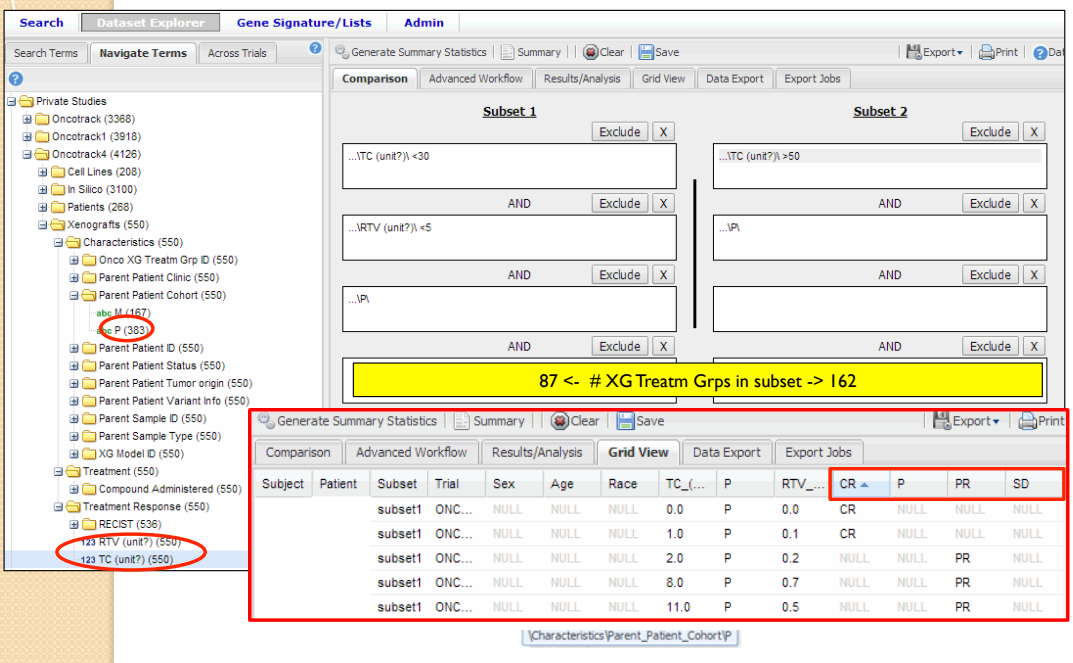
Patient ID	XG Treatm Grp
597-CM-M	597-CM-M-MF-01-04-XEN-01-BI_mTOR_FR
597-CM-M	597-CM-M-MF-01-04-XEN-01-BI836842
597-CM-M	597-CM-M-MF-01-04-XEN-01-Regorafenib
597-CM-M	597-CM-M-MF-01-04-XEN-01-Selumetinib
633-CB-P	633-CB-P-TF-01-04-XEN-01-BI_mTOR_FR
...	...

Cascading queries between distinct studies with 1xN relationship between them
- **Via joint variable – patient id**

Requirements
- **Export list with distinct Patient IDs related to a selection in either study**
- **Import that list as a selection criterium in the other study**

Interim solution
- **Lists of interest as characteristics in complementary study**

Flexible querying to hand over to analysis tooling



4.6 U-BIOPRED

Contributors: Kai Sun (ICL) and Florian Guitton (ICL)

4.6.1 Project description

U-BIOPRED is a multiple stakeholder, EU-IMI funded, severe asthma research project, comprising longitudinal clinical (adult and pediatric cohorts) studies and associated animal, *in vitro* and *in silico* model translational studies, using multi-omics technologies, attempting to produce novel classifiers better describing asthma disease heterogeneity.

4.6.2 What data types have been curated?

Samples collected from both human subjects and models are profiled using the following omics technologies:

- Transcriptomics - Affymetrix Gene Arrays
- Proteomics - MSE-based Label-free technology
- Lipidomics - ESI, HPLC-MS, GS-MS, LC-QTOF and MRM, depending on the lipid subset profiled
- Breathomics - eNOSE, GC/MS, NMR
- Genetics - Affymetrix SNP Arrays

N.B: Somalogic refers to the SOMAscan platform[8].

[8] http://www.somalogic.com/Products-Services/SOMAscan.aspx

The data types and volume of data curated and loaded in the U-BIOPRED eTRIKS/tranSMART instance to date, is shown on the table below (**Table 3**):

Technology	Datasets	Number of Subjects/Sample	Number of features	Total Features
Transcriptomics	Blood	202	54,675	11,044,350
	Bronchial Brushings	145	54,675	7,927,875
	Bronchial Biopsies	136	54,675	7,435,800
	Nasal Brushings	78	54,675	4,264,650
	Sputum	139	54,675	7,599,825
				38,272,500
Proteomics	Serum Somalogic	598	1,129	675,142
	Serum MS using MS^e technology[9]	131	130	17,030
	Sputum MS using MS^e technology[9]	99	3,231	319,869
				1,012,041
Lipidomics	Sputum MS using MS^e technology[9]	128	15,622	1,999,616
	Urine Eicosanoids	584	13	7,592
	Sputum Eicosanoids	319	13	4,147
				2,011,355
Beathomics	Adult eNOSE	106	190	20,140
	Adult GC/MS	64	7,036	450,304
	Paediatric eNOSE	106	190	20,140
	Paediatric GC/MS	64	7,036	450,304
				940,888
Clinical Data	Adult	617	2,388	1,473,396
	Paediatric	260	1,685	438,100
				1,911,496
Total				**88,296,560**

Table 3: Overview on data types and volume of curated data loaded in the U-BIOPRED eTRIKS/tranSMART instance

4.6.3 What methods/algorithms and/or pipelines have been developed/used?

4.6.3.1 Scripts

Custom scripts have been developed for pre-processing the clinical and lab results datasets, using R, Perl and Python languages and are publicly available on Github[10].

[9] http://www.medscape.com/viewarticle/814689_3

4.6.3.2 Annotation files

For all omics datasets custom annotation files were created in order for them to be "loadable" into tranSMART. This was achieved by leveraging the standard gene expression (GEX) platform annotation file structure. Furthermore, for proteomics datasets, all UniProt IDs were converted to EntrezGene IDs, thus enabling the filtering by pathway functionality available in several Advanced Workflows.

4.6.3.3 Loading scripts

Standard Kettle Scripts were used for all loading purposes.

4.6.3.4 Data Quality Control (QC)

Multiple data sources and diverse data types make the entire data management process highly error prone. Thus, to ensure provision of high quality data to the consortium, apart from automated rule-based check included in the pre-processing scripts, a data QC process was implemented.

4.6.3.5 Data QC process

- All data labelled "PROVISIONAL DATASET" was prepared and loaded into tranSMART as a secure study, which no consortium member except for platform admin and curators can access.
- Approximately 10 "testers" who are domain experts and omics platform experts were identified.
- A group comprised of the above testers was created and provided full access to the PROVISIONAL DATASET.
- An issue-tracking sheet was set up using Smartsheet (smartsheet.com) by BIOSCI consulting, that only the testers and data management team could have access.
- Testers were given two weeks, in which they were assigned to investigate the dataset and perform analytical queries.
- All identified issues were reported on the issue tracker.
- Once the issues had been resolved the amended dataset was loaded into tranSMART and all consortium members were granted access.

4.6.4 What problems have been encountered?

The only issue encountered was data upload time. Currently, the entire clinical and omics datasets require ca. 60 hours for complete upload. This was addressed by only performing data upload on weekends and by disabling access to tranSMART by rerouting the URL (transmart.doc.ic.ac.uk) to a "maintenance page" and informing all consortium members via e-mail, 2-3 days in advance.

[10] https://github.com/pandis83/UBIO_Scripts

UBIOPRED tranSMART April Upload (Fri 22/04/2016)

The finalised baseline clinical data are labelled as "Adult Cohort (Beta Testing – April 2016)" and "Paediatric Cohort (Beta Testing – April 2016)" on tranSMART.

- Atopy data – the latest atopy datasets provided by Graham are loaded onto tranSMART.
- Haematology and biochemistry data – the tranSMART overwrite rules are applied to the latest CROM download and the updated data are loaded onto tranSMART.
- TLC predicted and TLC actual predicted percentage – the values are recalculated based on the formula provided by Peter Sterk and the updated values are uploaded onto tranSMART
- Exacerbation - the variable "Exacerbation Number" under "Subject History" is removed as agreed.
- In all clinical variables, "Sarbutamol"s are replaced with "Salbutamol"s.

UBIOPRED tranSMART May Upload (Mon 16/05/2016)

The updates include updates on all finalised data we have received by 15th May, as previously agreed:

Baseline OMICS dataset:

- Update Philips GC-MS dataset (adult and paediatric cohorts)
- Update eNose dataset (adult and paediatric cohorts)
- Update SOTON serum proteomics dataset (adult cohort)
- Update SOTON sputum proteomics dataset (adult cohort)
- Update Human Protein Atlas dataset (adult cohort)
- Update Boehringer Ingelheim Cytokines and Chemokines dataset (adult cohort)
- Update Genentech Cytokines and Periostin dataset (adult cohort)
- Update Karolinska hsCRP dataset (adult cohort)
- Update Luminex dataset (adult cohort)
- Remove SOTON lipidomics sputum dataset from tranSMART, as the data is out-dated.

Longitudinal OMICS dataset:

- Update Philips GC-MS dataset (adult and paediatric cohorts)
- Update eNose dataset (adult and paediatric cohorts)
- Update Boehringer Ingelheim Cytokines and Chemokines dataset (adult cohort)
- Update Genentech Cytokines and Periostin dataset (adult cohort)
- Update Karolinska hsCRP dataset (adult cohort)
- Update Luminex dataset (adult cohort)
- Update Karolinska Eicosanoid Lipidomics dataset (adult cohort)

UBIOPRED tranSMART July Upload (Mon 04/07/2016)

The following datasets have been updated and are available on tranSMART for beta testing:

- The paediatric longitudinal eNOSE data is now on tranSMART ready for beta testing.

UBIOPRED tranSMART July Upload (Mon 25/07/2016)

The following datasets have been updated and are available on tranSMART for beta testing:

- Lung biopsy remodelling data (adult, broncoscopy visit) is now available on tranSMART under Adult Cohort/Clinical Data/Lung Biopsy Immunopathology/Broncoscopy Visit. The data is ready for beta testing.
- Blood handprint clustering data (adult) is now available on tranSMART under Adult Cohort/Subject Clusters. The data is ready for beta testing.
- Luminex Serum data (adult, baseline and longitudinal) is now updated and ready for beta testing.

UBIOPRED tranSMART August Upload (Mon 22/08/2016)

The following datasets have been updated and are available on tranSMART for beta testing:

- Karolinska Eicosanoid Lipidomics dataset (adult, baseline)
- Boehringer Ingelheim Cytokines and Chemokines (adult, baseline and longitudinal)
- Biopsy remodelling data (adult)

UBIOPRED tranSMART October Upload (Mon 10/10/2016)

The datasets below have been updated onto tranSMART and can be accessed from the "Beta_Testing" folder in the platform and are now ready for beta-testing:

- Platform - Drugomics Data:
  - Karolinska drug level data (adult baseline and longitudinal) are uploaded and ready for beta-testing.

- Platform - Lipidomics Data:
  - SOTON lipidomics data (adult baseline plasma and sputum data) are uploaded and ready for beta-testing.
  - Krakow Eicosanoid lipidomics data (adult longitudinal and paediatric baseline) are uploaded and ready for beta-testing.

- Clinical - Longitudinal Clinical Data:
  - All clinical data that were included in the "Beta Testing - Jun 2015 upload" are re-uploaded onto tranSMART.
  - Adult longitudinal 1.1 clinical data: tranSMART overwrite rules are applied to haematology and biochemistry data. TLC predicted and TLC actual predicted percentage values are recalculated and updated.

- Clinical - Exacerbation Data:
  - Adult and Paediatrics exacerbation data (screening, longitudinal 1, longitudinal 1.1, exacerbation day 1, telepost contact) are re-curated from the latest Nubilaria download (April 2016 download) and uploaded onto tranSMART. The "Start date month" and "Start date year" variables can be found under each exacerbation event.

- Clinical - Clinical Clustering:
  - TV clusters are updated as requested.

4.7 Bibliography:

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., … Soboleva, A. (2011). NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Research*, *39*(Database issue), D1005–10. doi:10.1093/nar/gkq1184

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., … Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Research*, *35*(Database issue), D760–5. doi:10.1093/nar/gkl887

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991–5. doi:10.1093/nar/gks1193