European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.11 – 4th progress report on Data Curation

Due date of deliverable: September 2016

Actual submission date: January 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:	D4.11
Deliverable title:	4th progress report on Data Curation
Deliverable version:	
Due date of deliverable:	October 1st 2016
Actual submission date:	January 2017
Leader:	Reinhard Schneider, Manfred Hendlich
Editors:	
Authors:	Adriano Barbosa, Wei Gu
Reviewers:	Chris Marshall; David Henderson
Participating beneficiaries:	UL, CNRS
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:	2
Nature:	Document
Version:	v1
Draft/Final:	Final
No of pages (including cover):	11
Keywords:	Curation, Data

4th Progress Report on Data Curation

1 Abstract

Data curation has been provided so far for several Innovative Medicines Initiative (IMI) projects as well as for the European Translational Information and Knowledge Management services (eTRIKS) Public server.

A set of IMI projects have so far collaborated with the eTRIKS consortium on the level of data curation, namely UBIOPRED, OncoTrack, RA-MAP, ABIRISK, APPROACH and AETIONOMY, which eTRIKS support has been already reported in the D4.9 deliverable

In this document, we report and update on the projects for which eTRIKS has provided support for curation since October 2015.

2 Curation training and documentation

No training has been conducted during the reported period. However, future trainings are scheduled.

3 Overview on eTRIKS supported projects

3.1 APPROACH

Wei Gu (UL) and Andreas Tielmann (Merck)

3.1.1 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.

3.1.2 What data types have been curated?

Until the time of this deliverable, eTRIKS has been working on the curation of two public 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.

3.2 AETIONOMY

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

3.2.1 Updates on curated datasets

Using the same curation description as described in the 3rd curation report, we supported the curation of the following datasets:

- PD Methylation data from partner UniKlinik Bonn (Germany);
- AD Cytokine data from partner UniKlinik Bonn (Germany);
- PD transcriptome data from partner ICM (France);
- AD transcriptome data from partner IDIBAPS (AD Screening cohort, Spain);
- PD transcriptoma data from partner BI (Germany);

3.2.2 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.

3.2.3 What problems have been encountered?

No major problems. We have used the same strategy delineated at the 3rd curation report.

3.3 RA-MAP

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

3.3.1 Updates on curated datasets

RA-MAP's TACERA study has increased both in terms of size and data types. Flow cytometry, X-ray scores, proteomics (SOMAscan) and metabolomics (1D and 2D NMR and Metabolon) have been curated and uploaded to tranSMART.

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

The methods applied to RA-MAP are the same described in previous reports

3.3.3 What problems have been encountered?

Problems were found to upload metabolomics to tranSMART. Scripts available at the tranSMART foundation github were incomplete and contained bugs. Errors were traced and solved. These new scripts are available at the tranSMART foundation github.

3.4 ABIRISK

Contributors: Wei Gu (UL)

No major update since the 3rd progress report release.

3.5 OncoTrack

Contributors: Adriano Barbosa (UL); Wei Gu (UL); David Henderson (Bayer); Gino Marchetti (CNRS); Anthony Rowe (JNJ); Venkata Satagopam (UL); Emmanuel Van der Stuyft (JNJ)

No major update since the 3rd progress report release.

3.5.1 What (new) problems have been encountered?

Due to restrictions on data use resulting from the wording of the patient's IC and delays to completion of the Data Processing Agreement, it was necessary to close the OncoTrack tranSMART instance in March 2015. The DPA has been signed by all parties and reactivation of the OncoTrack instance was possible in November 2016.

3.6 U-BIOPRED

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

3.6.1 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.

3.6.2 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)

3.6.3 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.

3.6.4 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.

3.6.5 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)

3.6.6 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 eTRIKS Public server

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

4.1 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 a 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 4.2 Description section deals with Gene Expression Omnibus (GEO)[1] database from where the public studies are retrieved. Section 4.3 deals with the types of data that are fetched and curated from the GEO database. The 4.4 section gives the details methods, Algorithm and pipeline used. In 4.5 we discuss problem encountered during data curation and upload.

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

For this release, a new data set is made available by eTRIKS Public server team consisting of 20 Asthma related studies with 1573 samples. The complete overview is given in Table 1.

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

Table 1: **Asthma GEO studies loaded to the eTRIKS/tranSMART Server**. The column Study ID indicates the name of the GEO selected. Domain column specifies the disease domain of the study: Asthma. Gene Expression Platform shows the IDs of the platforms used for gene expression study. Samples correspond to the number of samples deposited on GEO for each study. Variables column shows the number of clinical variables collected for each sample of one study. Data points refers to the total number of clinical values collected for each study.

Study ID	Domain	Gene Expression Platform	Samples	Clinical Variables	Data points
GSE45111	Asthma	GPL6104	47	6	282
GSE59339	Asthma	GPL8490	62	6	372
GSE41861	Asthma	GPL570	138	5	690
GSE41862	Asthma	GPL570	116	5	580
GSE41863	Asthma	GPL570	56	5	280
GSE67472	Asthma	GPL16311	105	5	525
GSE63383	Asthma	GPL6244	24	3	72
GSE35643	Asthma	GPL6244	12	14	168
GSE56553	Asthma	GPL13534	96	14	1344
GSE46171	Asthma	GPL16981	91	5	455
GSE52074	Asthma	GPL13534	18	7	126
GSE43696	Asthma	GPL6480	108	4	432
GSE31773	Asthma	GPL570	40	8	320
GSE22324	Asthma	GPL6104	200	6	1200
GSE65163	Asthma	GPL13534	72	13	936
GSE65204	Asthma	GPL14550	69	11	759
GSE65205	Asthma	GPL13534	141	13	1833
GSE45847	Asthma	GPL16979	42	18	756
GSE61225	Asthma	GPL19169	74	7	518
GSE40240	Asthma	GPL6244	28	30	840
GSE44037	Asthma	GPL13158	34	6	204
Total	**1**	**12**	**1573**	**191**	**12692**

4.3 Type of curated data

The curated data consist of clinical data along with the gene expression data for the Asthma studies.

4.4 Methods/Algorithms/Pipelines used

A pipeline is built using python 2.7, for extracting Asthma studies from GEO database, parsing the data and then generating standard format files. These format files are uploaded to tranSMART using Kettle Pentaho[3] ETL scripts[4].

In parallel to the above pipeline, during the report period, eTRIKS has received the demand to load pathway information to tranSMART. Facing the issues regarding

[3] http://community.pentaho.com/projects/data-integration/
[4] https://git.etriks.org/transmart-dse-etl/tree/master/DSE/Kettle/Kettle-ETL

licensing of KEGG[5] pathways, eTRIKS developed a script that permits users to load REACTOME [6] pathways instead. This script can be requested to eTRIKS by interested users.

4.5 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 many 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.

4.6 Bibliography:

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

[5] http://www.genome.jp/kegg/
[6] http://www.reactome.org/