IMI1 Final Project Report Public Summary

Project Acronym: ETRIKS

Project Title: Delivering European Translational Information & Knowledge Management Services

Grant Agreement: 115446
Project Duration: 01/10/2012 - 30/09/2018

1. Executive summary

Translational Research (TR), through the analysis of clinical and molecular data, provides new insights into disease progression, biomarker discovery, patient stratification and disease classification. TR intends to reduce the attrition of investigational new drugs during clinical development and the timelines associated with clinical programs. TR projects depend upon Knowledge Management (KM) capabilities and services that provide primary and secondary study data to project investigators.

Establishing a single TR KM platform and services will promote data and process harmonization across IMI projects, as well as other Public Private Partnerships. For individual projects, data/process harmonization will reduce operating costs, accelerate information system implementation and facilitate data utilization across the partners.

The eTRIKS consortium delivered a core TR KM software platform, TR analytics applications and a wide variety of value added best practices that impacted over sixty client projects during the course of the collaboration well exceeding the consortium's key goal and performance indicator of support for 40 projects. The consortium's assets are available, by and large, under open licenses and the application of best practices continues through the work of the eTRIKS commercial spinoff ***Information Technology for Translational Medicine***, the ***eTRIKS Data Sciences Network*** and the many adopters of eTRIKS' technologies.

1.1. Project rationale and overall objectives of the project

Overall Objectives

1. **Service:** Deploy and host the eTRIKS platform based upon the tranSMART Data Warehouse and provide training, support and consultation activities to all IMI project partners
2. **Platform**: Develop a tranSMART-based eTRIKS platform as a sustainable, interoperable, collaborative, re-usable, scalable and open source/open license TR KM technology stack supported by effective analytics methods and tools https://portal.etriks.org/portal/
3. **Content:** Establish eTRIKS as an unique European TR data resource supporting cross-organisation TR studies, including clinical studies and pre-clinical studies, omics data analysis for biomarker discovery and validation, genetics and NGS studies.
4. **Community:** Promote an active eTRIKS-centric international TR analytics & informatics community through active stakeholder engagement and by disseminating tools and expertise worldwide.

The eTRIKS platform will be further developed during the extension period (6) to integrate the feature-rich tranSMART version 17.1. Promotional activities will be conducted in support of the new integrated platform. The period 6 activities are described in detail in section 4.2, and new deliverables are also recorded below in section 1.2.

1.2. Overall deliverables of the project

Deliverable	Status
eTRIKS Knowledge Management Platform v4	Delivered
Public server with enhanced tranSMART 16.2	Delivered
Docker downloadable instance of eTRIKS KM Platform	Delivered
40 Public-Private Partnership projects supported	Delivered – 61 projects supported
180 publicly available studies with curated translational research data sets	187 delivered
eTRIKS Website and Portal to eTRIKS services	Delivered
Sustainable eTRIKS Network plan	Delivered
Standards Starter Pack v1.1	Delivered
Code of Practice on Secondary Use of Medical Research Data	Delivered
"Value of Data" Play-Decide utility to enhance patient engagement	Delivered
Data Catalogue, a searchable proof of concept repository for study metadata	Delivered
eTRIKS Training programme for users, curators and administrators	Delivered
eTRIKS Labs extensions to tranSMART - SmartR, HiDome, Disease Networks and Disease Maps, SNF, WGCNA, XNAT interface, eAE, Galaxy interface	Delivered
Outreach and dissemination	BioTransR conference a success

With the approval of the no-cost extension the following deliverables have been included for Period 6:

- D2.9 Integrated version of eTRIKS/tranSMART platform v5.0
- D5.9 Period 6 report
- D5.10 Final report
- D6.7 eTRIKS labs tested in two high profile use cases
- D6.8 eTRIKS book
- D6.9 Publication of Use Cases

1.3. Summary of progress versus plan since last period

Fewer partners participated during period six with Biosci Consulting, Imperial College and the EISBM operating through the use of their residual project funding. Some in kind/gratis effort was provided by select Pharma partners including Roche, Bayer and Pfizer. As a result, only work packages two (Software Development), five (Program Management) and six (Outreach) contributed deliverables during Period 6. However, the goals of period 6 were met by this reduced set of partners.

eTRIKS/tranSMART platform v5.0

eTRIKS platform v5.0, was developed to provide interoperability between the core eTRIKS platform, now based on tranSMART version 17.1, and the eTRIKS Analytic Environment (eAE). Additionally, a new user interface, called ***Borderline***, that was created for eTRIKSv5 as tranSMART 17.1 was not backwards compatible with the legacy eTRIKSv4/tranSMART 16.2 interface. The Borderline interface and high performance eAE (described more fully below) analytics integration met deliverable 2.9.

eTRIKS Book

The eTRIKS-penned book provides clinicians, translational researchers and other interested readers instruction in, and justification for, the use of exploratory clinical and biomarker data to achieve biomedical insights. The content is based on the real world experience of eTRIKS personnel during the course of the consortium. A decision on publisher is pending with a publication date planned for 2019. The pre-publication book meets eTRIKS deliverable 6.8.

eTRIKS High Profile Use Cases

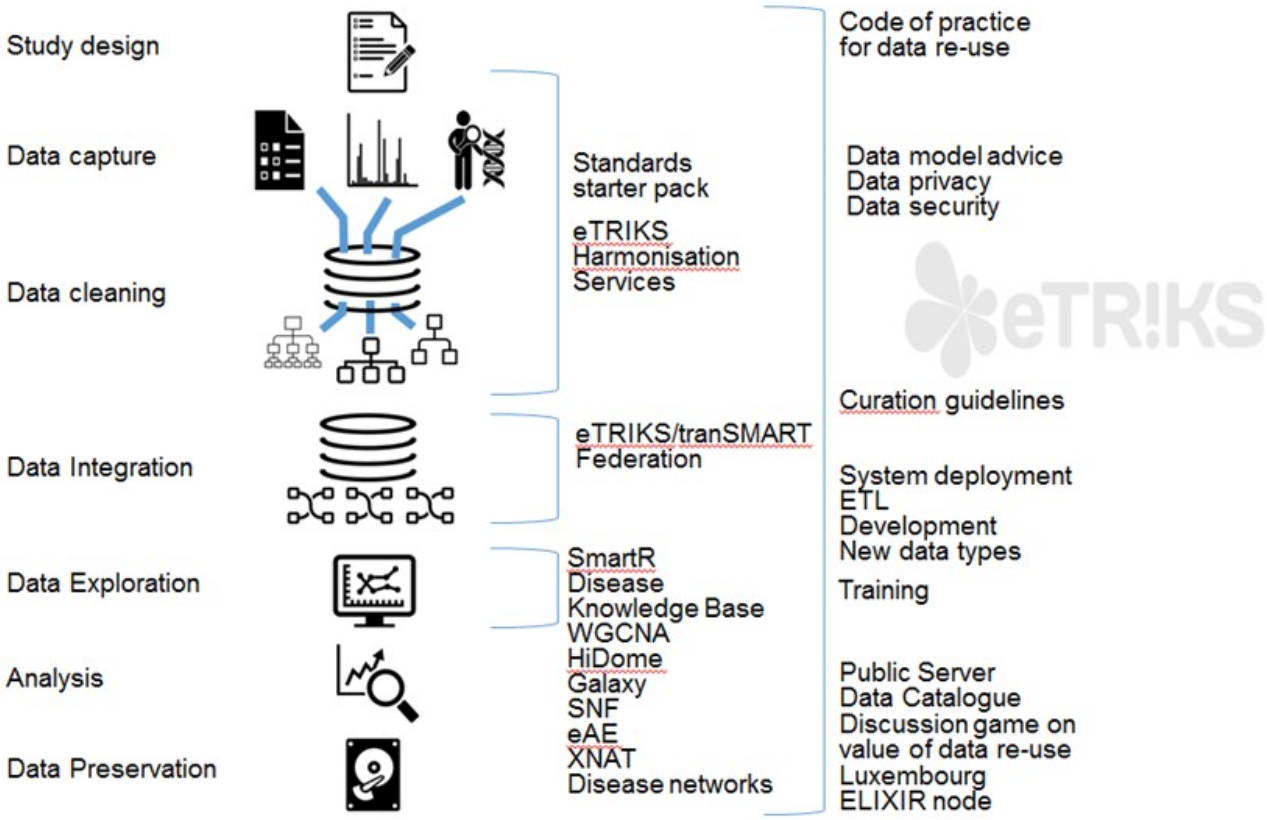
Deliverable 6.7 provides a detailed description of the application of eTRIKS' products and services with respect to the UBioPred and BioVacSafe consortia as exemplars of the potential of eTRIKS to enable translational data management and analysis. These exemplars were published and presented to key opinion leaders within the field of translational medicine at a "think tank" style meeting sponsored by eTRIKS and held during June 2018.

1.4. Significant achievements since last report

- eTRIKS platform upgraded (to version 5) incorporating tranSMART version 17.1.
- Achievement of an integrated environment where the tools of eTRIKS can be used together to support both data access and data exploration
- eTRIKS Analytical Environment integrated with eTRIKS Platform v5 .A new user interface, Borderline, created for eTRIKS version 5.
- eTRIKS Book written and pending publication in 2019.
- Exemplars of eTRIKS products and services were documented and presented.
- Think Tank meeting for Translational Research key opinion leaders convened by eTRIKS.
- Disease maps community developed and expanding

1.5. Scientific and technical results/foregrounds of the project

Figure 1: *eTRIKS assets and best practices and best practices aligned with the translational research study process*



The following foreground has been created or extended by eTRIKS over the course of the collaboration:

eTRIKS Translational Research Information Platform

The eTRIKS translational research information platform is based on the open source ***tranSMART*** translational research data warehouse created by ***Johnson and Johnson (J&J)***, adopted by the ***IMI-UBiopred*** project prior to the start of eTRIKS and made open-source under the ***GNU Public License version 3*** by J&J in 2012. eTRIKS released five major platform versions during the course of the collaboration.

- **eTRIKS Platform Version 1.0** incorporated the well established open source ***relational database management system*** (RDBMS) ***PostgreSQL*** into the tranSMART software stack such that tranSMART users could choose between ***Oracle***, the commercial enterprise RDBMS upon which J&J originally built tranSMART (Oracle is also licensed by many of the Pharma partners) and the fully open source PostgreSQL-associated deployment. The PostgreSQL integration was started by the University of Michigan prior to eTRIKS' launch. However, eTRIKS Work Package (WP) two personnel assumed integration responsibilities eventually providing the majority of effort required to complete this highly complicated and time consuming implementation. Twenty percent of WP2's development time was dedicated to the PostgreSQL integration and eTRIKS version 1 became a critical achievement with regard to the subsequent distribution of the platform across the many academic and public private partnerships that eTRIKS was to eventually support. eTRIKS 1.0 was co-released by the ***tranSMART Foundation*** (now the ***I2B2/Transmart Foundation***, https://transmartfoundation.org/) as tranSMART v1.1 to further promote exposure of the platform.

- **eTRIKS Platform Version 2.0** integrated several community tranSMART code branches and included early implementations of project workspaces, longitudinal data support and cross-study analytics. eTRIKS v2 wasmade to interoperate with the open source ***Galaxy*** data analysis environment (https://usegalaxy.org/). The integration of the community code was performed predominantly by eTRIKS WP2 personnel in direct collaboration with the tranSMART Foundation, which co-released the integrated system as tranSMARTv1.2.

- **eTRIKS Platform Version 3.0** was an incremental release to increase the platform's stability and robustness. Version 3.0 consolidated updates across multiple tranSMART v1.2.x minor releases that resolved various

operational defects. eTRIKS WP2 personnel continued to contribute substantially to these updates, in concert with the tranSMART Foundation, while also shifting focus to eTRIKS-specialized development efforts such as the eTRIKS Analysis Environment (eAE) and eTRIKS Harmonization Service (eHS) (eAE and eHS are detailed below). The tranSMART Foundation shifted its release nomenclature from *"major.minor"* notation to a calendar based notation comprised of year/semi-year (i.e. 2016.1st Half/2nd Half). eTRIKS version 3 was co-released as version 16.1 by the tranSMART Foundation.

- **eTRIKS Platform Version 4.0** added important analytic extensions including the integration of the open source image management system ***XNAT*** for medical image support, the **SmartR** interactive visual analytics interface invented by WP4, the **Hi Dome** extension enabling cohort selection/comparison based on high dimensional data types and the incorporation of support for **Genome Wide Association Study** (GWAS) summary statistics. eTRIKS 4.0 was built on the tranSMART Foundation 16.2 release which was the final stable version release of tranSMART v1.2/eTRIKS 2.0.

- **eTRIKS Platform Version 5.0** was developed upon tranSMART release 17.1 and made directly interoperable with the eAE. Additionally, a new interface, called ***Borderline***, was created for version 5, a necessity as the pre-tranSMART 17.1 interface was never made compatible with tranSMART v17.1. tranSMART v17.1 was created by a collaborative venture involving the tranSMART Foundation and four Pharma companies (Sanofi, Pfizer, Roche and AbbVie) with eTRIKS members Sanofi/Pfizer/Roche providing over 80% of the project funding. The application was developed by the software company ***The Hyve*** (https://thehyve.nl/) based in Utrecht, Netherlands. tranSMART 17.1 added more comprehensive support for longitudinal studies, cross-study analysis and support for RNA Sequencing. The eTRIKS period 6 extension was predicated upon the potential of developing a fully consolidated eTRIKS platform based on tranSMART v17.1. Therefore, the tranSMART v17.1 development funding provided to the Hyve was allowed as direct spend against Sanofi's, Pfizer's and Roche's eTRIKS commitment.

The development and deployment of the eTRIKS platform was clearly a substantial undertaking involving many consortium participants across WPs 1 (deployment and hosting), 2 (software development), 4 (analytics and data curation/mapping) and 6 (software requirements, testing). Requirements were also informed from work packages 3 (data standards) and 7 (ethical data use).

eTRIKS Public Platform: eTRIKS deployed and hosted a publicly-accessible eTRIKS Translational Research Platform (https://public.etriks.org/transmart/datasetExplorer) exposing clinical studies curated to eTRIKS' standards across a wide breadth of disease areas, eventually numbering roughly 200 studies by the end of period 5. The public platform met a key performance goal, set forth in the Description of Work, of providing open access to high quality curated data associated with 180 public domain studies. Additionally, the Public Platform served as a demonstration and training environment for investigators interested in using the eTRIKS platform. Moreover, deploying new eTRIKS Platform versions to the public server marked the completion of eTRIKS' software quality management process indicating readiness for client use and accreditation as a formal eTRIKS deliverable.

eTRIKS Labs

The eTRIKS Labs (https://www.etriks.org/etriks_labs/) concept arose out of a desire to centrally brand and distribute eTRIKS' many application and analysis method contributions that were developed and deployed supplementary to the eTRIKS translational research platform. All eTRIKS Labs are provided open-license to the research community. The eTRIKS Labs are comprised of the following assets:

- **eTRIKS Analysis Environment (eAE):** A high performance compute grid scheduler developed at the ***Data Sciences Institute at Imperial College London*** and deployed to the Institute's compute cluster. The environment enables investigators, and their applications, to launch compute jobs directly against the scheduler. Jobs can be launched using integrated ***Jupyter Notebooks***. The eAE software deployment is decoupled from specific high performance compute cluster implementations and, as such, the eAE can be installed and operated, in principle, on any cluster configuration.

- **eTRIKS Harmonization Service (eHS):** A system that facilitates the challenging manual data transformation and mapping process to the eTRIKS platform. The highly variable nature of data collected from clinical studies complicates its incorporation into structured data warehouses such as the eTRIKS platform. The eHS provides

an user interface to accelerate the configuration of clinical data collections and provides certain automated mapping capabilities. The eHS transformations are based on the industry standard ***Clinical Data Interchange Consortium Standards*** (**CDISC,** https://www.cdisc.org/ **)**. Pangaea enterprises (https://www.pangaeaenterprises.co.uk/) is a new company intending to apply artificial intelligence methods to facilitate clinical data curation. Pangaea is using the EHS as a core product to support their current curation services and as a starting point for developing and applying their automated curation extensions. Pangaea emerged in Period 6 providing a welcome element of sustainability for the eHS.

- **Hi Dome:** An application that allows eTRIKS platform users to select cohorts using values of high dimensional datasets, such as gene expression, and to determine statistically significant differences between the cohorts based on these high dimensional results (e.g. as significant differences in the expression of one or more genes between the cohorts). Hi Dome is a natural extension of tranSMART's base clinical data capabilities to high dimensional study results.

- **Disease Knowledge Base:** A semantic query application for molecular pathways created using the open source **Neo4J** (https://neo4j.com/) graph database engine leveraging the natural fit of semantic/graph databases to support the data structure of molecular networks. Molecular pathways structured as triple store relations can be searched using Neo4J's ***Cypher*** query language and presented visually with the corresponding information associated with each entity in the molecular network.

- **Disease Maps:** eTRIKS extended disease pathway maps related to Asthma and Parkinson's Disease working directly with information known a priori and data generated by client projects. Additionally, supplemental tools were created to accelerate the modeling of these disease maps from underlying disease-associated data.

- **Similarity Network Fusion (SNF):** An ***R-Shiny*** (https://shiny.rstudio.com/) application that was developed to provide an operational user interface for this novel computational method for genomic data integration (developed by Wang et al., in the lab of Anna Goldenberg (http://compbio.cs.toronto.edu/SNF/SNF/Software.html)). SNF constructs patient similarity networks based on a diversity of associated data types and, in a second step, iteratively integrates the individual patient networks until the algorithm converges to a final fused network representing the population.

- **Weighted Gene Co-Expression Network Analysis (WGCNA):** An R-Shiny application was developed to provide an operational interface for performing correlation network gene clustering analysis using the method implemented and published by Langfelder and Horvath (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/).

- **Patient Input Platform:** eTRIKS created a discussion game framework, based on the open license ***Play Decide*** platform (https://playdecide.eu/), to assist patients and legislators in navigating the risks and benefits of consenting their individual health data, or the data of their constituents, to promote biomedical research. A series of game cards that present questions to drive open discussion with respect to the topic of medical data reuse were created. With the help of a facilitator, groups of people work together to formulate/reassess opinions of the risks and potential of health data reuse. Applied in multiple sessions with patients, legislators and medical professionals the favorability of medical data sharing was consistently raised among these key stakeholders as a result of these sessions.

eTRIKS Standards Starter Pack

The selection and application of consistent data standards is critical for enabling high quality data review and analysis. Moreover, consistent data standards facilitate meta analyses across studies and increase opportunities for data reuse. The Standards Starter Pack documents the best practices for optimizing the quality and usability of exploratory medical data loaded to the eTRIKS platform. Tailored for project leaders and data managers, the resource provides a comprehensive review of pertinent biomedical data standards including guidance as to which standards platforms are best suited for specific research plans. The Standards Starter Pack documents were made available for all IMI projects to promote consistency in data handling and to raise awareness of the potential advantages of applying consistent standards across translational research projects. The Starter Pack was the basis for eTRIKS project consulting with respect to standards implementation. Multiple extended versions of the Standards Starter Pack were released by WP3 during the collaboration.

eTRIKS Code of Practice on Secondary Reuse of Medical Research Data

The Code of Practice provided multi-partner, multinational scientific research projects with urgently needed practical guidance compliant with (at the time of the original writing) applicable laws (i.e. the Eu Directive). The Code of Practice was adopted by the IMI, for all new projects, as the base level guideline for the design of ethical practices and policies regarding the appropriate use of patient data. The Code of Practice was developed and initially released by WP7 during period two to promote compliance with the Eu Directive. As such, the pertinence of the Code or Practice was lessened once the ***General Data Protection Regulation*** (GDPR) became law in May of 2018 (during the eTRIKS extension period). The Code of Practice was the basis of eTRIKS consulting with regard to ethical data use, however, WP7 members also became highly knowledgeable with respect to the GDPR statutes, pre-adoption of the GDPR into law, to assist client projects with the anticipated transition from the Directive to the GDPR. The ***BBMRI-ERIC GDPR Code of Conduct*** (http://code-of-conduct-for-health-research.eu/), currently in draft form, will fully replace the eTRIKS Code of Practice.

Data Catalogue

eTRIKS created the first broadly applicable ***Data Catalogue*** (https://datacatalog.elixir-luxembourg.org/) for datasets associated with projects from both IMI-1 and IMI-2 as well as other published sources. The Data Catalogue providesa searchable metadata repository that encompasses a wealth of cross-project study information allowing investigators to quickly find and assess datasets pertinent for their research endeavors. The Data Catalogue is a web-based product implemented using the Open Source ***CKAN*** (https://ckan.org/) data portal software and affords end users the opportunity to interactively search and display study metadata and summary information across the managed study collection.

Materials Transfer Agreement/Confidential Disclosure Agreement Templates

Material Transfer Agreements (MTAs) are contracts that define policies and responsibilities regarding oversight of the exchange and use of ***intellectual property*** (IP) between two or more parties. MTAs were generally necessary in order for eTRIKS to provide comprehensive services to client projects (research data being the pertinent IP for eTRIKS engagements). The MTAs between eTRIKS/***IMI-ABI-Risk*** and eTRIKS/***IMI-Oncotrack*** each required approximately two years of negotiations to close as all of the individual partners across the contracting consortia were required to authorize the MTAs as signatories (The ABI-Risk consortium alone required over 40 parties to enter into and complete negotiations). The experience of contracting across these large public private partnerships was codified into a MTA template containing the basic collection of terms and clauses pertinent to materials transfer. The template should accelerate the negotiations and subsequent execution of future MTAs. A similar template was created for ***Confidential Disclosure Agreements*** (CDAs) which were pertinent to eTRIKS project engagements. Although CDAs for IMI projects can be brokered via the project coordinators, serving as the signatory on behalf of all consortium participants, greatly reducing the time and effort to close a CDA relative to a MTA, nevertheless, the availability of the CDA template should further ease the start up time and cost of multi-project engagements.

eTRIKS Training Materials

eTRIKS personnel provided many training sessions during the course of the collaboration as part of WP6 outreach and promotion efforts. Topics codified into training programs and materials include:

- Platform installation and support
- Introductory guide for new platform users
- In depth training for advanced platform users
- eTRIKS reporting (defect management and support services)
- Building new interfaces with the tranSMART API
- Data Privacy and Reuse
- Application of Data Standards
- Introductory Data Curation and Database Mapping
- Advanced Data Curation and Database Mapping
- Electronic Case Report Form design

The training materials that were developed for these sessions were not licensed for open use as these assets were considered important to the eTRIKS sustainability program. The training materials comprise the only proprietary foreground maintained by the consortium.

eTRIKS Website (https://www.etriks.org/) and domain name *eTRIKS.org*
eTRIKS appropriated the ***eTRIKS.org*** domain and created and maintained an extensive website to provision the public foreground and promote the efforts of the consortium.

Post Consortium Asset Allocation
To foster sustainability eTRIKS grant recipients assumed responsibility for each foreground asset (excepting as noted by "*") at the end of Period 5 as follows.

Imperial College London (Yike Guo)
- eTRIKS Platform
- eHS
- EAE
- eTRIKS Website

U of Luxembourg (Reinhard Schneider)
- Public platform w/ curated studies
- Client platforms hosted at the Computing Center for The National Institute of Nuclear and Particle Physics in Lyons, France
- SmartR
- Data Catalog
- Service Portal
- R-Shiny SNF
- R-Shiny WGCNA

European Institute for Systems Biology and Medicine (Charles Auffray)
- Disease Knowledge Base
- Disease Maps

Biosci Consulting (Scott Wagers)
- Patient Input Plan
- Training Materials
- MTA/CDA Templates

U of Oxford (Susanna Sansone)
- Standards Starter Pack

* Hi Dome is maintained by J&J, the Code of Practice is maintained on the eTRIKS website but is deprecated with respect to the BBMRI-ERIC Code of Conduct.

1.6. Potential impact and main dissemination activities and exploitation of results

eTRIKS deliverables were comprised mainly of foreground assets and best practices associated with the application and use of the foreground. The primary ***Key Performance Indicator*** (KPI) for eTRIKS was the application, in whole or in part, of these deliverables to advantage client projects and research organizations. The expressed consortium goal of supporting 40 clients was attained late in period 4 with 61 client projects supported by the end of period 5. Six clients (***ABIRISK***, ***U-BIOPRED, PRECISEADS, PreDiCT-TB, ONCOTRACK, APPROACH)*** were provided with comprehensive support by eTRIKS that included eTRIKS platform provisioning, data curation and loading, custom development and analytics, if needed, and standards/ethical/training consultation. Thirty-four projects were provided with enabling support which included select (one or more) applications of foreground assets and/or best practices. Cases in which client projects deployed and maintained the eTRIKS Platform without the involvement of eTRIKS personnel were counted as enabled projects as were projects which utilized eTRIKS consulting to create their data management plans. The majority of enabled projects, drawn by the economic benefits of the PostgreSQL fully open source option, used the eTRIKS platform. Sixteen additional projects were associated with four organizations that used the eTRIKS platform for their own clients. In many cases these organizations were not at

liberty to disclose the identities of their client projects but instead provided the count of client projects for which the organization supported instances of the eTRIKS platform. Two projects nearing their termination dates, ***SAFE-T*** and ***COPD Map***, used the eTRIKS platform to consolidate study data for persistent access by the Pharma partners following the end of the project. Both SAFE-T and COPD Map contracted the eTRIKS platform services via the eTRIKS sustaining entity ***Information Technology for Translational Medicine*** (ITTM), a commercial eTRIKS spinoff founded by Reinhard Schneider at the University of Luxembourg and contracted through the ***eTRIKS Data Sciences Network*** (eDSN). Two projects, ***BIOVACSAFE*** and ***OPTIMIZE***, received resourcing for custom application development.

Figure 2: *Supported Consortia and Projects as of ~month 52*

eTRIKS accomplished much of this outreach and brand recognition through direct promotion and communication by WP6, and other consortium personnel, to key stakeholders leveraging IMI connections, digital marketing and via participation in many conference-level events. The following outreach engagements are especially noteworthy.

BiotransR Conference (May-2017, Barcelona)

The BioTransR conference, organised by eTRIKS, was conceived with the intent to bring together data scientists and translational researchers to introduce leading edge translational data and analysis tools. These tools were promoted as opportunities to speed translational research, enhance confidence in translational analysis results and highlight the use of data to expand the breadth of translational research opportunities. The conference held a series of general discussions regarding the landscape of life sciences IT applications and infrastructures. Importantly, the general application discussions were followed by talks that emphasized case scenarios pertinent to the translational researcher and how the tools introduced in the earlier discussions have benefitted the research outcomes of such cases. eTRIKS applications and best practices were, of course, featured. However, introduction to the eTRIKS portfolio was coupled with the experience of almost five years of refinement and real world application.

European Union Parliamentary Event, Play Decide Multi-Stakeholder Discussion Sponsored by eTRIKS (October-2016, Brussels)

The European Union Parliamentary Event assembled a large number of patients, researchers, policy makers and members of the Eu parliament. The Secretary of State for Social Fraud and Privacy, Philippe De Backer, was in attendance. The Patient Input Plan's Play Decide Platform was used to drive discussions aimed at exposing and understanding the various perspectives regarding digital technologies that are producing, managing and analyzing large scale health data collections derived from individual patients. Specifically, the assembly was tasked with addressing the, presumed, contrasting imperatives of protecting the privacy of patients with respect to the disclosure of their health information versus the potential of these data to drive medical breakthroughs through consented reuse of these data. Following the session the majority of participants indicated that research data should be used beyond the project for which it was collected (~80%) and that research data can be shared in certain instances where there is adequate personal protection (100%). Most participants responded in favor of compulsory data sharing for research purposes (~65%). Opinions were mixed with respect to personal ownership of an individual's own health data and the requirement of obtaining an explicit consent for each proposed use of health data.

eTRIKS Labs Launch (October-2015, Amsterdam)

The eTRIKS-Labs brand was launched at an eTRIKS-specific community meeting which introduced the eTRIKS-Labs web portal and the available labs including the eAE, eHS and SmartR. The meeting immediately followed the tranSMART Foundation's annual meeting with the intent of maximizing the exposure to potential clients and technologists capable of using the labs to enable their own research clients. The eTRIKS labs brand consolidated the diverse assets across eTRIKS that were not conceptually straightforward or technically sensible to integrate into the core eTRIKS technology platform. The intent in consolidation was to cohesively promote diverse program elements, such as the molecular pathways databases, R-Shiny analytics applications and Patient Input Plan to enhance awareness and adoption as well as to foster a sense of a shared consortium vision among the Lab's developers.

Training Courses Over Periods 2 to 5

Over the course of the collaboration eTRIKS Platform training was provided to hundreds of translational researchers. Generalized training was provided to groups of investigators representing various projects and interests. A tailored training approach was used when all attendees were affiliated with a single project. The tailored training made use of project-specific data sets that had been curated in advance by WP4 and project-specific use cases that had been thoroughly tested and vetted in advance of the training session(s). Where applied, the project-specific training approach was, as would be expected, especially effective in expediting use at the cost of substantial planning time and commitment by eTRIKS personnel. Generally, project-specific training was reserved for the comprehensively-supported projects.

eTRIKS created information and technology applications to speed clinical research studies. The tools enabled, and continue to enable, a diverse set of scientific research projects aimed at advancing the understanding, and subsequent development, of therapeutic interventions for major diseases. In this manner, eTRIKS benefits the health of Eu citizens and the competitiveness of European research community. The continued support of eTRIKS assets by the consortium's academic partners, commercial software service companies such as ITTM and the Hyve, ongoing collaborative health ventures such as the eTRIKS Data Sciences Network and tranSMART software developers will promote the sustainability and continued dissemination of eTRIKS' work products.

1.7. Lessons learned and further opportunities for research

Developing application infrastructure and corresponding best practices to the magnitude accomplished by eTRIKS required a highly focussed effort by a large group of diversely skilled individuals. Although such an effort might have been achievable by a sizeable software company motivated by the opportunity to serve translational researchers, the sheer diversity of information management needs across prospective clients would have challenged any company trying to create a broadly applicable solution with an equivalent comprehensive set of best practices. Indeed, several commercial entities build and/or provide services for translational information management

solutions. Such companies often provide specialized services (ITTM, Rancho Biosciences, Clarivate Analytics) or focus on the enhancement of existing products for individual clients (the Hyve) and become an integral part of the translational sciences business network. Certain commercial attempts at building enterprise translational solutions have been discontinued while others have gained traction through committed development and thoughtful marketing (Perkin Elmer).

The preferences of prospective clients are, of course, critically important. Academic clients, either acting as a single project team or within the context of a Public Private Partnership, were the key customer groups targeted by eTRIKS. The inherent willingness of academic customers to partner with strongly-affiliated academic consortia, such as eTRIKS, and make use of the corresponding open license applications that eTRIKS was obligated to produce, was an advantage for eTRIKS in converting academic prospects to clients.

Therefore, the goal of meeting eTRIKS' client-based key performance indicator was probably best achieved through a public private partnership. However, the focus on client projects came at the cost of direct benefit to Pharma partners. Indeed, only three of the ten Pharma partners adopted tranSMART for internal use and specific Pharma partner needs became subservient to the needs of client projects. The primary, and largely indirect, benefit to eTRIKS' commercial participants were derived from their simultaneous participation in eTRIKS *and* eTRIKS-supported client projects.

Figure 3: *Top 20 Public Private Partnerships supported by eTRIKS and their respective eTRIKS partner affiliates as a qualitative measure of indirect value for Pharma. Comprehensively-supported projects highlighted in red.*

	Sanofi	Roche	AstraZeneca	Janssen	Bayer	Merck	Lilly	Lundbeck	Pfizer	GSK	
UBioPred		x	x	x						x	4
ABIRISK	x				x	x			x	x	5
Oncotrack		x	x		x	x	x		x		6
PreciseADS	x				x		x				3
Predict-TB	x			x						x	3
Approach						x				x	2
COPD MAP			x						x	x	3
SAFE-T	x	x	x		x		x		x	x	**7**
BioVacSafe	x									x	2
Matura		x		x					x		3
RA-MAP		x	x	x					x	x	5
Aetionomy	x										1
Quic-Concept	x	x	x			x	x		x	x	**7**
BioAster	x										1
EMIF		x		x		x			x	x	5
ND4BB	x		x	x						x	4
Masterplans		x									1
PSORT				x					x	x	3
EU-AIMS		x		x			x		x		4
Genomics England		x	x							x	3
Comprehensive	3	2	2	2	3	3	2	0	2	4	
Total	**9**	**10**	8	8	4	5	5	0	**10**	**13**	

The following lessons learned may be of value for consideration by future consortia.

Collaborative Use of Data

eTRIKS intended to provide data services to client projects using a centralized application hosting environment based at CCIN2P. An infrastructure plan was designed and implemented to support a substantial number of projects that were expected to generate large volumes of molecular biomarker data. Given certain statutes that cite regional limits with respect to the transfer and storage of individual-level medical data, coupled with the inertia experienced

with closing MTAs described above, the centralized infrastructure served only two client projects in addition to the eTRIKS Public Platform. The infrastructure was eventually well purposed to support the eTRiKS Labs, which included demonstration instances of many of the software applications. However, the creation of an European-centralized multi-tenant translational research information platform that would host IMI, and other European government-funded biomarker studies, was a critical goal that motivated several Pharmaceutical companies to participate as eTRIKS partners. The vision of a pan-European resource for translational biomarker data sets, akin to the U.S. NIH data commons, was an important driver for eTRIKS that was legally unattainable under the varied statutes of the member nations of the European Union. Additionally, the diversity of data sharing policies associated with the target translational research projects also impeded the aggregation of project data sets. Although eTRIKS was to become highly active in the promotion of data sharing via the ***eTRIKS Manifesto (arguing for an open health data ecosystem)***, Patient Input Plan and support for/participation in the ***FAIR (Findable, Accessible, Interoperable, Reusable) Data*** community, the inability to meaningfully aggregate data across client projects was a profound frustration for the consortium. New consortium efforts should consider this eTRIKS experience, during the drafting of their descriptions of work, to address:

1. The collection, management, use and dissemination of data that is foreground, and sideground, across the collaboration partners including policies for:
    a. Confidential disclosure
    b. Materials transfer
2. Licensing and management of data that is background.
3. A data management plan for use external to the collaboration, if pertinent, that details schedules and policies for licensing and distribution.
4. An evaluation of legal constraints and risk assessment that may impact the creation, transfer and storage of data that is foreground/sideground.
5. Consider assigning a data officer/office to implement consortium policies and adjudicate issues raised by stakeholders, especially for consortia responsible for highly diverse/complicated data domains.
6. Consider opportunities such as the ***ELIXIR Europe*** (https://www.elixir-europe.org/) program which places compute infrastructure nodes within Eu member states such that data and applications can, in principle, be aggregated/consolidated within national borders. Data resident on an Elixir node in one member state can then be federated to corresponding data/application environments hosted on the Elixir nodes of other member states once contractual and legal constraints are satisfied. The eTRIKS public platform is now hosted on an Elixir node supported by Reinhard Schneider's group at the University of Luxembourg in order to promote federated data sharing via eTRIKS platforms distributed across the Elixir network.

Diversity of Program Requirements and the Challenges of Software Reuse

eTRIKS planned to develop an explicit set of extended tranSMART capabilities including:

1. Fully-open source software stack (PostgreSQL integration)
2. Cross study cohort selection and analysis
3. Support for longitudinal study designs
4. Compliance validation/application auditing
5. Application Federation

As described prior, WP2 incorporated PostgreSQL to create a fully open source tranSMART option. Although a complicated and time consuming effort implemented by the WP2 team, the desired outcome was very straightforward to articulate and test as this change, implemented effectively, would be inconsequential with respect to the tranSMART end user experience. Moreover, the value prospect of the PostgreSQL integration, primarily one of

economic advantage and freedom of use, was singularly applicable and beneficial across all of eTRIKS' prospective clients.

By contrast, the remaining list of items are straightforward in concept, i.e. standardized representation of data parameters/values across studies and time-relevant data structures with corresponding search features. However, clinical studies typically collect hundreds to thousands of distinct data elements, the set of which can differ substantially across studies, even studies designed for a single disease area. Both calendar time and event-based relative timing (e.g. days post first visit) may be pertinent to analysis. Satisfying such exacting and study-specialized requirements mandate detailed specification and design and is unlikely to be feasible during the pre-project planning phase. Indeed, the diversity of possible cross study and longitudinal implementations challenged eTRIKS as initial enhancements provided by the tranSMART community for eTRIKS version 2.0 were unable to satisfy eTRIKS' clients. Implementations of these features driven by eTRIKS' partners for eTRIKS version 4.0 were still incomplete requiring an eTRIKS extension period to develop a robust and useable software deployment and solution. Moreover, it is likely unreasonable to expect that eTRIKS version 5 provides the definitive implementation of the cross study and longitudinal features as there is likely *no* definitive feasible implementation that will satisfy all specialized use cases of prospective projects.

However, the eTRIKS platform was able to satisfy several projects having cross study and longitudinal analysis needs through custom enhancements to the core eTRIKS platform and/or, more routinely, via creative data curation and database mapping by expert data scientists. With complex software use patterns there will be challenges to software reuse and opportunities for customization. In these instances, customizing a base solution, such as the eTRIKS platform, will offer many future projects an accelerated path towards a productive solution to specialized requirements. ITTM, the eSDN (and others) are positioned to assist clients with such extensions to the eTRIKS platform.

eTRIKS partners and clients requiring federation and clinical validation to be implemented in tranSMART were not accommodated. Clinical validation was determined not to be pertinent for most of eTRIKS' prospective clients and federation was deprioritized once the consolidated multi-tenant project environment was deemed impractical and the ramifications of legal barriers to data transfer were more fully understood.

It is important for new translational sciences consortia to understand that translational data management expectations may vary substantially across studies and across data consumers. Aspects of data consumption that are critical to individual partners should be represented, and thereby committed, directly in the initial statement of work provided that these requirements are understood at the outset of a nascient consortium.

More generally, partner expectations and priorities with regard to project outcomes may also vary substantially, the extent of which may not become evident until the project is underway. The academic partners clearly approach the consortium with specialized interests, expertise and planned outcomes. Academic teams cannot be regarded as outsourcing partners that can mold/modify their talents to fit shifting consortium directives, or changes in consortium direction. Rather, the expertise of the academic partners must be fully aligned with the goals of the consortium at launch. Changes to the remit of the consortium may disrupt this alignment and the consortium management must respond thoughtfully to such challenges realizing that academic partners may be neither willing nor able to accommodate substantive changes to the collaboration's initial goals.

Similarly, although most IMI Pharma partners have an extraordinarily large and diverse workforce, the likelihood, as experienced during the eTRIKS term, of exchanging in-kind resources allocated pre-project to accommodate project changes that require a different combination of expertise will be low. For example, companies that allocated IT/technical personnel were not easily able exchange these resources for legal and compliance personnel to staff a

work package, established very late in the pre-project negotiation stage, dedicated to the ethical use of human data.

A diversity of expectations across partners is to be expected. For eTRIKS this diversity was manifest in varied levels of interest across participants for system attributes such as validation, federation or aggregate study content. Challenging these goals by the realization of impracticality or diminished relevance, coupled with eTRIKS' key performance indicator prioritizing client needs over those of partners, led to decreased motivation, and subsequently lessened commitment, for some Pharma partners. Additionally, harmonization of activities can be disrupted when focus areas shift. For example, the availability of aggregate disease area study collections favors system biology methods development and application. The shift in focus from multi-tenancy to distributed project-specific deployments would be expected to decouple the core system from biological network analytic method development and application. Addressing such decoupling through harmonized branding, such as eTRIKS Labs, is a thoughtful response for conserving and valuing the analytical aspirations of specialized expert academic teams. However, the resolution, viewed at the level of the consortium, is less powerful given the reality that direct technical integration of the analytic software to the core system will not provide the envisioned benefit. The analytic tools continue development, serve many projects, but are used, published and valued more as individual rather than collective contributions, unavoidable under the conditions of changing goals and less satisfying as a collaborative outcome.

As best as possible, explicitly documenting and committing to project outcomes in the initial DoW will, or at least should, facilitate management decisions in the face of (inevitable?) changes in scope. However, for software, or software-associated projects, the reality of complex requirements and the time and risk involved in defining such requirements essentially necessitate their detailed specification to be a project, rather than pre-project, activity carrying the high risk of unanticipated changes in initial scope and expectations. If substantial changes are encountered expect the beneficiaries to continue to pursue their initial plans with limited ability to alter their consortia goals. As the commercial participants will also be unlikely to accommodate changes that necessitate an alternate distribution of staffing/expertise, expect the participation of commercial participants to diminish under such circumstances.

When formulating the eTRIKS call and providing guidance to respondents, the IMI and Pharma leads addressed the software requirements risk specifically by selecting, and electing to enhance, an existing open platform with IMI project precedence and experience. This highly prudent decision set the foundation for eTRIKS to achieve an incredible distribution of products and services. However, this solid pre-project planning for a complex software project could not fully protect the collaboration's ideal vision from the serious consequences of unexpected legal and cultural barriers to data consolidation. Six years later, this pre-project perspective of not assessing the data sharing risk, at least explicitly, will seem naive on the part of the eTRIKS partners. However, at the time that eTRIKS was first conceived (2008) and launched (2012) attempts at consolidating such comprehensive sets of patient health data across European national borders were few.

For stakeholders investigating the potential of large-scale collaborative software projects via the IMI, or similar agency, the following points are worth considering.

- Codify stakeholder desires as explicitly as possible in the Description of Work
    - Commit to certain outcomes (eTRIKS example: Priority of client support) and consider the value of these collaborative outcomes with respect to the desired outcomes of the individual partners.
- Expect changes that will challenge these desires once the project is underway
    - Consider risks and codify potential mitigation plans in the call and DoW (eTRIKS example: extend existing system (tranSMART) vs. custom build)
    - Go beyond the Risk Assessment "boilerplate" (personnel turnover, corporate reorganization, initial recruiting/onboarding of staff, etc. [these are of course valid and will become issues but mitigation

plans are as generalizable/superficial as the risks themselves]). Think critically to identify risks that are unique, or of specific project impact, that will require equally specialized mitigation plans. Such thinking may lead to mitigation plans that impact partner/personnel selection and shifting of project vision/goals.

- Partners may not be able to modify their intended contributions in the face of change
    - Academic partners will likely need to continue to work against their initial goals (eTRIKS example: Analytic method developers sourcing data from outside of the core platform and develop methods that are released as applications uncoupled from the core platform)
    - Some pharma partners may become disenfranchised with changes leading to diminished contributions. Accept this as a reality of long term complex collaborative software projects, however, some partners may realize unexpected opportunity/value in the face of change and contribute beyond their initial intent.
    - Formulate resolutions to maintain priority to committed high value outcomes (eTRIKS example: Retain focus on client support at the expense of accommodating individual interests to limit the perceived fracturing of the effort).
    - Find unorthodox ways of reinforcing collaborative attitudes (eTRIKS examples: muster independent efforts together and promote as a consolidated brand (eTRIKS Labs), group products and services into multidisciplinary work streams with shared goals (eHS development + standards + analytic development)
- Is a large scale long term committed collaboration the best model for the intended outcomes?
    - Smaller engagements will be more manageable with more limited results. l
    - If you are simply building software, does the novelty of the software merit a collaborative research model/investment. If so, should the commercial entities invest on their own, how will academic expertise propel the engagement?
- Circumstances may, or likely will, alter the initial vision of the project. Highly valuable, possibly unexpected, outcomes can be accomplished under changing conditions with persistence, consistent values, creative thinking and an appreciation for the interests of fellow partners.

The next logical steps following the end of the eTRIKS collaboration would be to pursue data strategies to support large scale translational data collections, i.e. a major opportunity that was pragmatically unattainable by eTRIKS. The eTRIKS platform has proven to be an excellent solution for individual, or small sets, of translational studies. In these cases the burden of data curation and mapping is both necessary and tractable while the inherent search and analysis capabilities may be fit for purpose, or made fit for purpose through small to medium scale software enhancement. However, leveraging large collections of data requires a different data management strategy and technology infrastructure. One approach would be to index the raw (or lightly processed) files comprising the data collection to enable search without costly pre-processing/data curation. Content identified as useful would be imported (following pertinent approval/consent activities) to data sciences packages for subsequent transformation and analysis. In essence, managing a large collection of data would be enabled by a model commonly referred to as a ***Data Lake*** as opposed to the ***Data Warehouse*** approach of the eTRIKS/tranSMART platform. However, follow-on efforts would need to address the legal and cultural data sharing challenges described earlier.