Advancing clinical research by semantically interconnecting aggregated medical data information in a secure context

Athos Antoniades[1,2], Aristos Aristodimou[1,2], Christos Georgousopoulos[3],Nikolaus Forgó[3], Ann Gledson[4], Panagiotis Hasapis [3], Caroline Vandeleur[6], Konstantinos Perakis[7], Ratnesh Sahay[8], Muntazir Mehdi[9], Christiana A. Demetriou[9], Marie-Pierre F. Strippoli[6], Vasiliki Giotaki[10], Myrto Ioannidi[10], David Tian[5], Federica Tozzi [1,2], John Kean[5] and Constantinos Pattichis[1]

[1]University of Cyprus, Nicosia, Cyprus, Email: athos@cs.ucy.ac.cy
[2]Stremble Ventures LTD, Limassol, Cyprus, Email: aristos.aristodimou@stremble.com
[3]INTRASOFT International, Luxembourg, Email: christos.georgousopoulos@intrasoft-intl.com
[4]Gottfried Wilhelm Leibniz Universität Hannover, Hannover, Germany, Email: nikolaus.forgo@iri.uni-hannover.de
[5]University of Manchester, Manchester, United Kingdom, Email:jak@cs.man.ac.uk
[6]Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland, Email: Caroline.Vandeleur@chuv.ch
[7]Ubitech Ltd., Athens, Greece, Email: kperakis@ubitech.eu
[8]Insight Centre for Data Analytics, NUI Galway, Ireland, Email: ratnesh.sahay@deri.org
[9]Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus, Email: christianad@cing.ac.cy
[10]Zeincro Hellas S.A., Vrillisia, Greece, Email: vgiotaki@zeincro.com

Abstract*— Electronic Health Records (EHRs) contain an increasing wealth of medical information. When combined with molecular level data, they enhance the understanding of the underlying biological mechanisms of diseases, enabling the identification of key prognostic biomarkers to disease and treatment outcomes. However, the European healthcare information space is fragmented due to the lack of legal and technical standards, cost effective platforms, and sustainable business models. There is a clear need for a framework facilitating the efficient and homogenized access to anonymized distributed EHRs, merged from multiple data sources into a single data analysis space. However, the diversity and complexity of today's legal and ethical regulations imposed at both the national and European level make it difficult, risky and expensive to transfer data by sharing EHRs. This is more evident when EHRs contain data that can't be easily anonymized such as genetic, or data related to rare conditions. In this paper we present the outcomes of Linked2Safety, a project that proposes a solution to these problems by providing a semantically interconnected approach to sharing aggregate data in the form of data cubes. This approach eliminates the risks associated with sharing pseudoanonymized (and therefore still personal) data while enabling the multi-source, multi-type analysis of health data through a single web based secure access platform. The proposed approach is evaluated using real data from three independent data providers from different EU countries. In this paper, external evaluators are presented the outcomes of the real data deployments and are given the opportunity to experience and evaluate the system with simulated data.The Linked2Safety system is evaluated by external to the project Medical science analysts, Analytic methodology engineers and Data providers with respect to five specific dimensions of the system (analysis space, linked data space, usability of the system, legal and ethical issues, and value of the system) in this paper. For all five dimensions that were examined, the participants' perceptions were overwhelmingly positive.

Keywords*— Semantic Interoperability, Electronic Health Records, Personal Data Protection, Adverse Event Prediction, Genetic Analysis

1. INTRODUCTION

Recent advances in medical research disciplines have enabled the collection of medical data with unprecedented dimensionality and volume, especially with regards to molecular level data (-omics). Analyses of such data have already enabled a better understanding of the underlying biological mechanisms and have led to novel new personalized therapies for some diseases. With the advent of –omics technologies into the clinical setting as key diagnostic biomarkers for predisposition and prediction of disease and treatment outcomes we see these technologies making their way into many patients' records enabling the discovery of key medical knowledge relevant to pharmacovigilance with a genetic component. However, the cost of performing such studies is high, and the increased dimensionality caused by the testing of orders of magnitude more hypotheses also increases the negative effects of the multiple comparisons problem. A need therefore exists to perform analyses across data providers from multiple independent international institutions. The problem there lies into two distinct aspects, addressing legal/ethical requirements at both the national and international level, and addressing the technical challenges of correctly aligning data together for meaningful analyses to take place without inducing bias or errors in the outcomes. Hence, the European healthcare information space is fragmented due to the lack of legal and technical standards, cost effective platforms, and sustainable business models.

This paper introduces a novel approach that has been developed and tested attempting to overcome the aforementioned fragmentation barriers by the Linked2Safety system. The system is capitalizing upon the latest technologies in order to realize a scalable, technical infrastructure, for the efficient, homogenized access and the effective utilization of the increasing wealth of medical information contained in the Electronic Health Records (EHRs) and Electronic Data Capture (EDC) systems deployed and maintained at regional and/or national level across Europe. The solution that is presented was developed and evaluated under realistic conditions in the context of the Linked2Safety project co-funded by the European Commission.

2. Background

The potential gains in efficiency and effectiveness for primary care when obtained by rapid and secure access to patient healthcare data in electronic form are widely recognized today across the EU. Providing an interoperability infrastructure for EHRs is on the agenda of many regional, pan-European and international eHealth initiatives [1,2,3], while about half of the member states are currently working on national eHealth infrastructures.

Towards this end, the semantic approaches to promote interoperability among standard-compliant information systems, e.g. reference ontologies and mediation, have proven to be able to have significant potential as regards the integration of information from distributed EHR databases. The semantic interoperability of patient data between EHRs and medical research can transform today's process of drug discovery, development and commercialization, enable faster access for patients to effective new medications, provide improved patient outcomes, improve medication security and signal detection, and provide a key foundation for targeted personalized medicines [2,4,5,6]. Furthermore, eHealth research during the last few years does not focus on healing, but rather proactively acting and keeping citizens healthy. In this sense, research at pan-European and international level highlights the importance of adverse event prevention in the healthcare domain, integrating heterogeneous data-sets from various clinical centers.

Along these lines, Linked2Safety proposes to build the next-generation, semantically-interlinked, secure medical and clinical information space in the enlarged Europe that will allow dynamically discovering, fruitfully combining and easily accessing medical resources and information contained in spatially distributed EHRs. Moreover it will leverage the reuse of EHRs in clinical research, towards the early detection of potential patient safety issues, based on genetic data analysis, extraction of bio-markers associated with an identified type of an adverse event, and advanced epidemiological research. It also aims to support sound decision making, towards the effective organization and execution of clinical trials, allowing healthcare professionals and medical scientists to easily submit their own query and get homogenized access to high-quality medical data.

3. Legal and Ethical Issues

Within projects like Linked2Safety, patients' personal health data is an important source of information. However, these data are – with very good reasons, as data protection is a key factor for the development of successful solutions – intensely protected. It is therefore necessary to thoroughly study the corresponding national and European legislative framework, and to establish specific legal and ethical requirements for the platform that is developed.

Many of the basic data protection principles in Europe are enshrined in article 6 of Directive 95/46/EC. For example, data may not be kept in a form which permits identification of data subjects for longer than is necessary for the purposes for which the data were collected, or for which they are further processed. In addition, specific data security measures need to be implemented: According to Art.17 para. 1 of the European Data Protection Directive [7], the data controller as well as the data processor, must implement appropriate technical and organizational measures to protect the data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access and against all other unlawful forms of processing. At the same time, the data subject has several rights that must be respected, such as the right to be informed, the right of access, the right of rectification, erasure or blocking and the right to object.

From the ethical perspective, the processing of health data at EU level requires the consent of the data subject [8], though in certain cases and under specific conditions this might not be needed. For instance, the national laws of the clinical pilot partners of Linked2Safety in Greece [9], Cyprus [10] and Switzerland [11] declare that the use and processing of health data without the consent of the data subject is possible, if and only if, it is for research purposes, the data are properly anonymised, and analysis is done in aggregated level.

4. Medical Data in the study

Table I Medical Data per Data Provider

Study	Type of Study	Host Institution	Number of Subjects	Target Population
MASTOS	Breast Cancer Case-Control Study	CING	2286	Greek - Cypriot Adult Female Population
DIABETES TYPE II	Diabetes Type II	CING	1070	Greek - Cypriot Adult Population
CoLaus	Population based study	CHUV	6734	Lausanne, Switzerland aged 35-75 years
PsyCoLaus	Population based study	CHUV	3,719	Lausanne, Switzerland aged 35-66-years
Phase III	Non interventional multicenter phase	ZEINCRO	101	Greek Adult population

clinical trials	IV clinical trial on respiratory.			
Phase IV clinical trials	Non interventional multicenter phase IV clinical trial on cardio.	ZEINCRO	3,125	Greek Adult population

4.1. Large Scale Genome Wide Association Study (GWAS) Epidemiological Data (CHUV)

The Linked2Safety project used aggregated data from the CoLaus and PsyCoLaus studies which were conducted in a population-based sample in Lausanne, Switzerland. Lausanne is the 5th largest city in Switzerland, localized in the French speaking part. First, the CoLaus study [12], based on a sample of 6,734 individuals randomly selected from the population registry of 35-75 year-old residents of the city of Lausanne, involved a comprehensive assessment of cardiovascular risk factors and collected DNA and plasma samples for the study of genetic variants and biomarkers. Participation was 43% [12]. Second, all 35 to 66-year old subjects of the CoLaus sample (n=5,535), were invited to participate in the psychiatric study from which 67% accepted (n=3,719) [13]. Ninety-two percent of them were Caucasians. The final sample for Linked2Safety used the data on subjects that had completed both the somatic and the psychiatric exams. The gender distribution of the sample (47.1% males) did not differ from that of the source population in the same age range and the mean age for the overall sample was 50.9 (s.d. 8.8) years. The Institutional Ethics Committee of the University of Lausanne approved the CoLaus and subsequently the PsyCoLaus study. All participants signed a written informed consent after having received a detailed description of the goal and funding of the study.

4.2. Candidate Gene Studies (CING)

4.2.1. MASTOS

MASTOS (Greek for "Breast") is to date the largest breast cancer case-control study in Cyprus carried out between January 2004 and December 2006, by the Department of Electron Microscopy / Molecular Pathology at the Cyprus Institute of Neurology and Genetics (CING). The purpose of MASTOS was to investigate the genetic and non-genetic epidemiology of breast cancer in Greek Cypriot women.

The cases consisted of 1109 women, 40-70 years of age with a histologically confirmed diagnosis of primary breast cancer. The control group consisted of 1177 Cypriot women from the general population, who had received a negative mammography result.

Demographic and risk factor data were collected from both cases and controls with the use of a specially designed questionnaire, through a standardized interview. In addition, a blood sample was taken from each subject, which was subsequently used for DNA extraction and genetic sequencing analysis. Genetic analyses in MASTOS focused on candidate genes, mostly involved in DNA repair pathways.

MASTOS received approval from the Cyprus National Bioethics Committee and was funded by the Research Promotion Foundation of Cyprus and the Cyprus Institute of Neurology and Genetics.

4.2.2 DIABETES TYPE II

DIABETES TYPE II, is a case control research program with the purpose of investigating gene regions of high influence for Diabetes Type 2 (DT2) in the Cypriot population. The study was carried out by the Institute of Neurology and Genetics (CING) in collaboration with the University of Cyprus, Makarios Hospital, and the Hippocratic Cyprus Diabetes Association. The research was funded by the Research Promotion Foundation of Cyprus.

The study recruited 520 healthy control subjects who had a recent measurement of blood glucose within the normal range. In addition, 550 patients diagnosed with DT2 were recruited.

All subjects were given a questionnaire to collect epidemiological and medical data and were asked to provide a small blood sample which was used to extract DNA and investigate candidate genes which were previously associated with DT2 in other populations. Laboratory results were studied using statistical methods and data mining to draw conclusions regarding the influence of these gene regions for the development DT2 Cypriot population.

The study received approval from the Cyprus National Bioethics Committee

4.3. Clinical Trial Data (ZEINCRO)

ZEINCRO is a contract research organisation (CRO) specialist in Central and South-Eastern Europe. The services that ZEINCRO provides to pharmaceutical companies in terms of clinical trials include recruitment and selection of study sites along with clinical monitoring and safety reporting throughout the clinical trial duration. For Linked2Safety ZEINCRO acquired approval from one of its biggest client/pharmaceutical company to use datasets from 7 clinical trials (phase III and phase IV) oriented to cardio and respiratory drugs. Each study was described by a separate protocol with predefined objectives and endpoints, inclusion/exclusion criteria. Confidentiality was preserved at each stage of Linked2Safety and under all circumstances.

4.3.1. Phase III Data

Phase III studies are trials with the purpose of determining the short and long-term safety/efficacy balance of formulation(s) of the active ingredient, and of assessing its overall and relative therapeutic value. The pattern and profile of any frequent adverse reactions must be investigated and special features of the product must be explored. For Linked2Safety the total number of subjects used was 101. Each study followed an individual protocol but the common data collected included patient demographics, smoking habits, medical history and other medical conditions including treatments as well as the adverse events throughout the study duration. In addition blood and biochemical test results are collected every time they occur. Each of these studies focused on a specific respiratory drug (mainly asthma treatment) and the main objective was to determine the efficacy and safety profile of these drugs.

4.3.2. Phase IV Data

Trials in phase IV are carried out on the basis of the product characteristics on which the marketing authorization was granted and are normally in the form of post-marketing surveillance, or assessment of therapeutic value or treatment strategies. The Phase IV clinical trial data used in Linked2Safety involved 3125 subjects. Once again each study followed its individual protocol but common data were again collected; including patient demographics, smoking habits, medical history, laboratory findings and other medical conditions including treatments as well as the adverse events throughout the study duration. Each study focused on a specific cardiovascular drug (mainly hypertension, diabetes, prevention of atherothrombotic events, hypercholesterolemia treatment) and the main objective was to establish the therapeutic equivalence between the test and reference drug.

4.4. Dataset Overlap

Despite the different study designs, research questions, and research institutions, the datasets provided by the three clinical partners demonstrated significant overlap, with respect to phenotypic as well as genetic variables.

Genetic data was available from CHUV, in the context of a GWAS study (CoLaus), and from CING in the context of a candidate gene association study on Diabetes Type II. Given the broad spectrum of Single Nucleotide Polymorphisms (SNPs) investigated in the GWAS, most candidate SNPs from CING were also investigated by CHUV. These 11 SNPs, with details of their position and significance, are presented in Table II.

Table II Genetic Variable Overlap between CHUV and CING

SNP	Gene	Gene Name	Chromosome	Genomic Locus	Allele Variations			Summary
rs10811661	CDKN2A/B	Cyclin-Dependent Kinase Inhibitor 2A/B	9	9p21.3	CC	CT	TT	significant for type-2 diabetes risk
rs10923931	NOTCH2	Neurogenic locus notch homolog protein 2	1	1p13-p11	GG	GT	TT	susceptibility to type 2 diabetes, obesity
rs10946398	CDKAL1	CDK5 Regulatory Subunit-Associated Protein 1-Like 1	6	6p22.3	AA	AC	CC	susceptibility to type 2 diabetes
rs1801282	PPARG	Peroxisome Proliferator-Activated Receptor-Gamma	3	3p25	GG	CG	GG	associated with type 2 (1) diabetes and fat metabolism (metabolic syndrom)
rs4402960	IGF2BP2	Insulin-Like Growth Factor 2 mRNA-Binding Protein 2	3	3q27.2	GG	GT	TT	significant for type-2 diabetes risk
rs4607103	ADAMTS9-AS2	ADAMTS9 antisense RNA 2	3	3p14.1	TT	TC	CC	significant for type-2 diabetes risk, obesity
rs5015480	HHEX	Hematopoietically Expressed Homeobox	10	10q23.33	AA	AG	GG	significant for type-2 (1) diabetes risk
rs7901695	TCF7L2	Transcription Factor 7-Like 2	10	10q25.3	CC	CT	TT	significant for type-2 diabetes risk
rs7961581	TSPAN8	Tetraspanin 8	12	12q14.1-q21.1	TT	TC	CC	susceptibility to type 2 diabetes, obesity
rs8050136	FTO	Fat mass- and obesity-associated gene	16	16q12.2	TT	TA	AA	association with body mass index, obesity risk, and type 2 diabetes
rs864745	JAZF1	JAZF Zinc Finger 1	7	7p15.2-p15.1	CC	CT	TT	significant for type-2 diabetes risk

With respect to phenotypic and dichotomous variables, there was significant overlap between the datasets of all three clinical partners. Table III lists the phenotypic categorical variables that were common between at least two of the datasets, and demonstrates that the raw material on which to demonstrate Linked2Safety platform's ability to link data was available.

Table III Phenotypic Categorical Variable Overlap between the Clinical Partners

	Contains Variable			Range	Label
	CING	ZEINCRO	CHUV		
Gender	Yes	Yes	Yes	1	"Male"
				2	"Female"
Age at interview Age at recruitment	Yes	Yes	Yes	0	"<30 yrs"
				1	"30-39 yrs"

				2	“40-49 yrs”
				3	“50-59 yrs”
				4	“60-69 yrs”
				5	“≥70 yrs”
Height	Yes	Yes	Yes	0	“≤150 cm”
Adult body height				1	“151-160 cm”
				2	“161-170 cm”
				3	“171-180 cm”
				4	“>180 cm”
Weight	Yes	Yes	Yes	0	“≤50 kg”
Weight at interview				1	“51-60 kg”
Maximum weight				2	“61-70 kg”
				3	“71-80 kg”
				4	“81-90 kg”
				5	“>90 kg”
Body mass index	Yes	Yes	Yes	0	“Underweight (<18.5 kg/m^2)”
				1	“Normal (18.5-24 kg/m^2)”
				2	“Overweight (25-29 kg/m^2)”
				3	“Obese (≥30 kg/m^2)”
Systolic blood pressure	Yes	Yes,	Yes	0	“Normal (<120 mmHg)”
				1	“Borderline high (120-139 mmHg)”
				2	“High (≥140 mmHg)”
Diastolic blood pressure	Yes	Yes	Yes	0	“Normal (<80 mmHg)”
				1	“Borderline high (80-89 mmHg)”
				2	“High (≥90 mmHg)”
Blood glucose levels	Yes	Yes	Yes	0	“Normal (<110 mg/dl (6.1 mmol/l))”
				1	“Borderline high (≥110 and <126 mg/dl (6.1-7.0 mmol/l))”
				2	“High (≥126 mg/dl (7.0 mmol/l))”
HDL cholesterol	Yes	No	Yes	0	“Low (<40 mg/dl (1.3 mmol/l))”
				1	“Normal (40-59 mg/dl (1.3-1.5 mmol/l))”
				2	“High (≥60 mg/dl (1.6 mmol/l))”
LDL cholesterol	Yes	No	Yes	0	“Optimal (<100 mg/dl (2.6 mmol/l))”
				1	“Near or above optimal (100-129 mg/dl (2.6-3.3 mmol/l))”
				2	“Borderline high (130-159 mg/dl (3.4-4.0 mmol/l))”
				3	“High (≥160 mg/dl (4.1 mmol/l))”
Total cholesterol	Yes	No	Yes	0	“Desirable (<200 mg/dl (5.2 mmol/l))”
				1	“Borderline high (200-239 mg/dl (5.2-6.1 mmol/l))”
				2	“High (≥240 mg/dl (6.2 mmol/l))”
Smokers	Yes	Yes	Yes	0	“Never”
				1	“Past”
				2	“Current”
				3	“Unknown/NA”
Number of pregnancies	Yes	No	Yes	0	“None”
				1	“One”
				2	“Two”
				3	“Three”
				4	“Four and more”

Table IV lists the availability of various dichotomous variables by each data provider and as can be seen there is a significant overlap in most of them in at least two of the data providers. The only exception is in some of the adverse event variables, which were present only by CHUV.

Table IV Dichotomous Variables Overlap between the Clinical Partners

	Contains Variable		
	CING	ZEINCRO	CHUV
Cardiovascular			
Hypertension	Yes	Yes	Yes
Myocardial infarction	Yes	No	Yes
Coronary heart disease	Yes	No	Yes

Endocrine/Metabolic			
Dyslipidemia	No	Yes	Yes
Diabetes	Yes	Yes	Yes
Diabetes type II	Yes	No	Yes
Ophthalmological problems			
Glaucoma	Yes	Yes	No
Cataracts	Yes	Yes	No
Myopia	Yes	Yes	No
Neurologic/Psychiatric			
Parkinson's disease	No	Yes	Yes
Depression	No	Yes	Yes
Schizotypal personality disorder	No	Yes	Yes
Medications (Cardio)			
Clopidogrel (B01AC04)	No	Yes	Yes
Carvedilol (C07AG02)	No	Yes	Yes
Simvastatin (C10AA01)	Yes	Yes	Yes
Doxazosin (C02CA04)	No	Yes	Yes
Oxerutin (C05CA)	No	Yes	Yes
Metoprolol (C07AB02)	No	Yes	Yes
Bisoprolol (C07AB07)	No	Yes	Yes
Amlodipine (C08CA01)	Yes	Yes	Yes
Verapamil (C08DA01)	No	Yes	Yes
Diltiazem (C08DB01)	No	Yes	Yes
Enalapril (C09AA02)	Yes	Yes	Yes
Enalapril and diuretics (C09BA02)	No	No	Yes
Lisinopril (C09AA03)	No	Yes	Yes
Perindopril(C09AA04)	No	Yes	Yes
Ramipril (C09AA05)	No	Yes	Yes
Cilazapril (C09AA08)	No	Yes	Yes
Losartan (C09CA01)	Yes	Yes	Yes
Eprosartane (C09CA02)	No	Yes	Yes
Valsartan (C09CA03)	No	Yes	Yes
Irbesartan (C09CA04)	No	Yes	Yes
Candesartan (C09CA06)	Yes	Yes	Yes
Telmisartan (C09CA07)	Yes	Yes	Yes
Ezetimibe (C10AX09)	Yes	Yes	Yes
Medication (Psy)			
Clonazepam (N03AE01)	No	Yes	Yes
Alprazolam (N05BA12)	No	Yes	Yes
Zolpidem (N05CF02)	No	Yes	Yes
Paroxetine (N06AB05)	Yes	Yes	Yes
Sertraline (N06AB06)	Yes	Yes	Yes
Piracetam (N06BX03)	No	Yes	Yes
Adverse events			
Any adverse events	No	Yes	Yes
Weight loss	No	No	Yes
Extrapyramidal side effects	No	No	Yes
Headaches	No	Yes	Yes
Sexual symptoms	No	No	Yes
Sleep problems	No	No	Yes
Metabolic syndroms	No	No	Yes
Prolactin symptoms	No	No	Yes
Trembling	No	No	Yes
Hair loss	No	No	Yes

5. The Linked2Safety Approach

The vision of Linked2Safety, is to ensure and empower patients’ safety, supporting clinical and medical research and improving the quality of healthcare. This is accomplished by providing patients, healthcare professionals and pharmaceutical companies with an innovative interoperability framework, a sustainable business model, as well as a scalable technical infrastructure and platform for the efficient, homogenized access to and the effective utilization of the increasing wealth of medical information contained in the EHRs and EDC systems. This allows dynamically interconnecting distributed patients data with clinical research efforts, respecting patients’ anonymity, data ownership and privacy, as well as European and national legislation.

In order to achieve this, Linked2Safety produced an open and generic software reference architecture based on which a prototype platform is delivered to support the reuse of semantically interlinked, interoperable EHR and EDC information resources. The platform provides healthcare professionals, clinical researchers and experts from pharmaceutical companies a user-friendly, sophisticated, collaborative decision-making environment that enables the analysis of all the available data of the subjects, such as genetic, environmental and their medical history during a clinical trial leading to the identification of the phenotype and genotype factors that are associated with specific adverse events, and thus early detection of potential patients’ safety issues. It also enables subject selection for clinical trials through the seamless and standardized linking with heterogeneous EHR repositories, providing advice on the best design of clinical studies.

5.1. The Data Cube approach

Recognizing the importance of the data protection, data security and anonymisation of clinical data originating from different clinical data providers, particular attention was given on the regulatory and security related aspects of the Linked2Safety platform where these data are being preserved and processed. The platform innovatively adopts the concept of data-cubes for transforming the legacy medical data into a form that makes impossible a patient’s re-identification via reverse engineering methodologies. In brief this is achieved in a three step procedure.

5.1.1. Data Cubes

Initially the data are converted into many small multidimensional data-cubes. Data-cubes can be seen as multidimensional contingency tables that only contain aggregated data. In Figure 1, a 3D data-cube is illustrated for two SNPs and a disease variable.

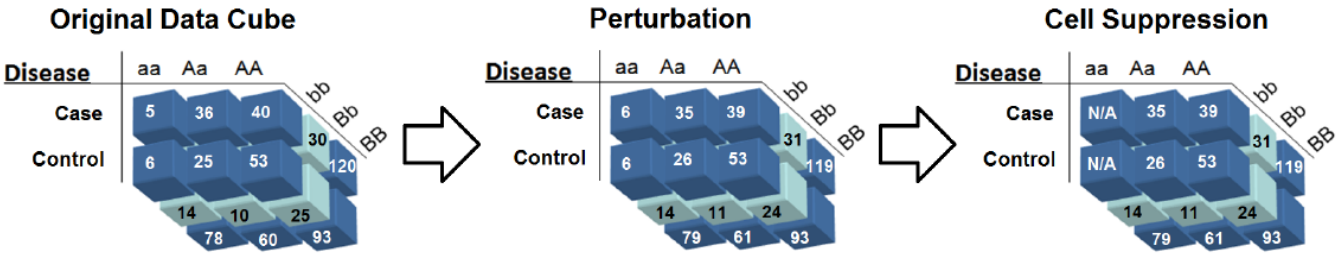


Figure 1Anonymisation of data utilising the data-cube approach

Each cell of the data-cube denotes the number of people that have certain characteristics. For example the top left cell indicates that there are 5 people with minor homozygous alleles (“aa”) in the first SNP and minor homozygous alleles (“bb”) in the second SNP and have the disease (“case”).

Perturbation

Once a data-cube is created, the values in each cell get perturbed. Perturbation is the addition of noise on the aggregated values of the data-cube. The noise added is in a specific range of values. In the illustrated example, the perturbation was in the range of minus one to one.

5.1.2. Cell Suppression

The final step is cell suppression. In this step, all cells with a value below a pre-specified threshold are removed from the data-cube. In the example shown, a threshold of 10 was used and resulted in removing two cells; this step is essential for preventing the re-identification of persons, since sensitive data are removed from the data-cubes. A method for identifying the maximum perturbation and cell suppression that can be used in a dataset without affecting the results of analyses is described in [14]

Therefore, all information maintained into EHR and EDC repositories which is accessible by the Linked2Safety platform is transformed into data-cube structures that include only aggregated counts of patients having specific characteristics instead of raw record-level information.

5.2. Semantically Linking & Querying Aggregated Medical Records

5.2.1. Linking

Once the data is converted into multidimensional data-cubes, the data-cubes are represented using the Resource Description Framework (RDF), which is a data model specific format. After storing RDF data-cubes in a context-aware fashion, the next challenge is to semantically link them to LOD (Linked Open Data) datasets that overlap with the domain of study and are thus targets for interlinking. However, given the growing diversity of the LOD datasets: 41 datasets on the LOD cloud are classified as specializing in the 'Life Sciences' domain and 70 Protocol And RDF Query Language (SPARQL) endpoints have been made available by the publishers; the initial step of finding relevant datasets that can potentially be linked to is a challenging task. Manually identifying which of these LOD datasets are potential targets for links with the local datasets of each clinical partner is a time consuming process.

Creating links is a challenging task for publishers. To address this challenge, a number of linking frameworks, such as Silk [15] and LIMES [16], have been proposed to help publishers link their local datasets to a remote LOD dataset through a specified SPARQL endpoint and are deployed as part of Linked2Safety platform.

5.2.2. Querying

Inspired by the publication of hundreds of Linked Datasets on the Web, researchers have been investigating federated querying techniques to enable access to this decentralized content. Query federation, aims to offer clients a single-point-of-access through which distributed data sources can be queried in unison. In the context of Linked Data, various optimized query federation engines have been proposed that can federate multiple SPARQL interfaces.

However, in the context of the Healthcare and Life Sciences (HCLS) domain, where data-integration is often vital, real-world datasets contain sensitive information: strict ownership is granted to individuals working in hospitals, research labs, clinical trial organizers, etc. Therefore, the legal and ethical concerns on (1) *preserving the anonymity of patients (or clinical subjects)*; and (2) *respecting data ownership through access control*; are key challenges faced by the data analytics community working within the HCLS domain.

The key challenges for federated querying are efficient source selection (determining which sources are (ir) relevant) and query planning (determining an efficient query execution strategy). Query-federation engines often apply source selection at the level of endpoints, whereas in a controlled environment, a user may only have access to certain information *within* an endpoint. Adding an access control layer to existing SPARQL query federation engines adds unique challenges: (1) source selection should be granular enough to enable effective access control, and (2) it should be policy-aware to avoid wasteful requests to unauthorized resources.

To facilitate this, SAFE, a SPARQL query federation engine that supports policy-based access to sensitive statistical data is proposed. SAFE is motivated by the needs of three clinical organizations in the context of Linked2Safety Project who wish to enable controlled federation over statistical clinical data -- such as data from clinical trials -- owned and hosted by multiple clinical sites, represented in the form of data cubes: multi-dimensional arrays of numeric data. The architecture of SAFE is given in Figure 2, and is detailed in [17] along with evaluation results. To better understand the functioning of SAFE, the following example is provided..

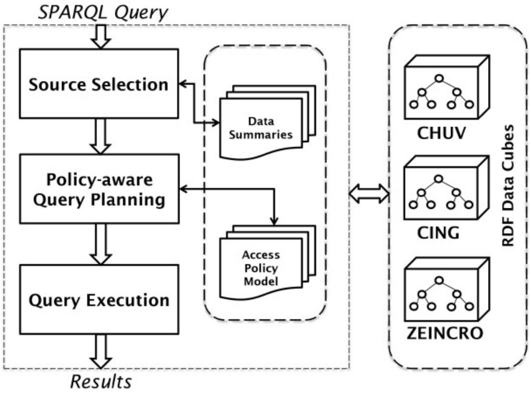
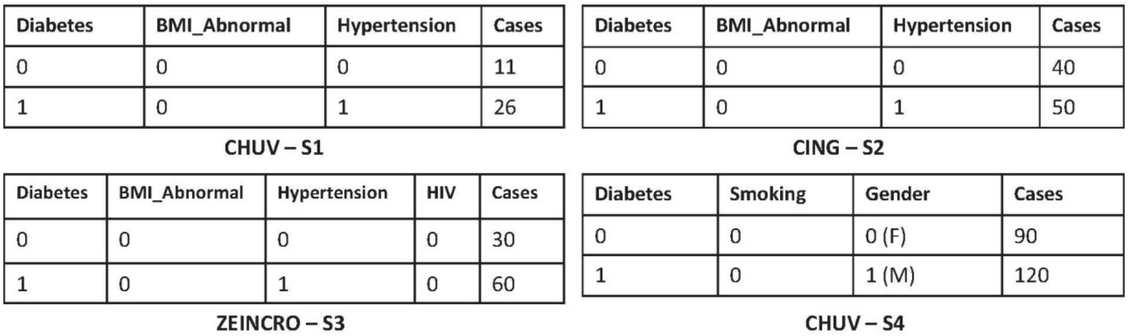


Figure 2 SAFE Architecture

Figure 3 shows four sample data cubes published by three different clinical sites. Each observation represents the total number of patients exhibiting a particular adverse event. For example, the *CHUV-S1* observations describe the total number of patients (in the *Cases* column) that exhibit a particular combination of three adverse events: *Diabetes*, (Abnormal) *BMI_Abnormal* (Body Mass Index) and/or *Hypertension*. The value *0* or *1* indicates if the condition is present or not. For example, the second row in *CHUV-S1* shows that there are 26 cases presenting with both *Diabetes* and *Hypertension* but without *BMI_Abnormal*.



Diabetes	BMI_Abnormal	Hypertension	Cases
0	0	0	11
1	0	1	26

CHUV – S1

Diabetes	BMI_Abnormal	Hypertension	Cases
0	0	0	40
1	0	1	50

CING – S2

Diabetes	BMI_Abnormal	Hypertension	HIV	Cases
0	0	0	0	30
1	0	1	0	60

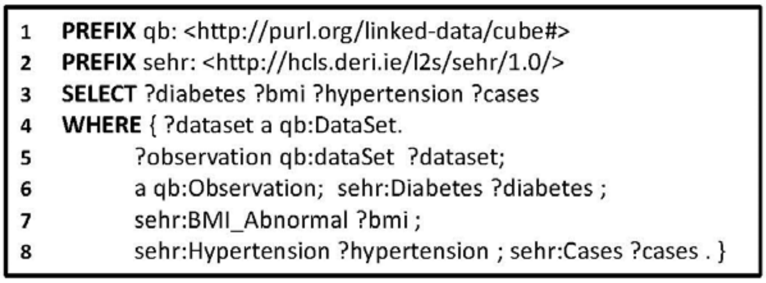
ZEINCRO – S3

Diabetes	Smoking	Gender	Cases
0	0	0 (F)	90
1	0	1 (M)	120

CHUV – S4

Figure 3 Example (2D) data cubes published by CHUV, CING and ZEINCRO

Once the data are published by clinical sites, they should be accessible to clinical researchers. Figure 4 shows a sample SPARQL query specifying subject-selection criteria, asking for the counts of cases that involve some combination of diabetes, abnormal BMI, and hypertension. An answer returned by the query, that is, number of cases, will play a major role in deciding the resources (number of subjects, location, etc.) required for conducting a clinical trial. However, answering such a query requires integrating RDF data cubes with three dimensions -- *Diabetes*, *Hypertension*, *BMI_Abnormal* -- and the respective counts originating from multiple clinical sites.

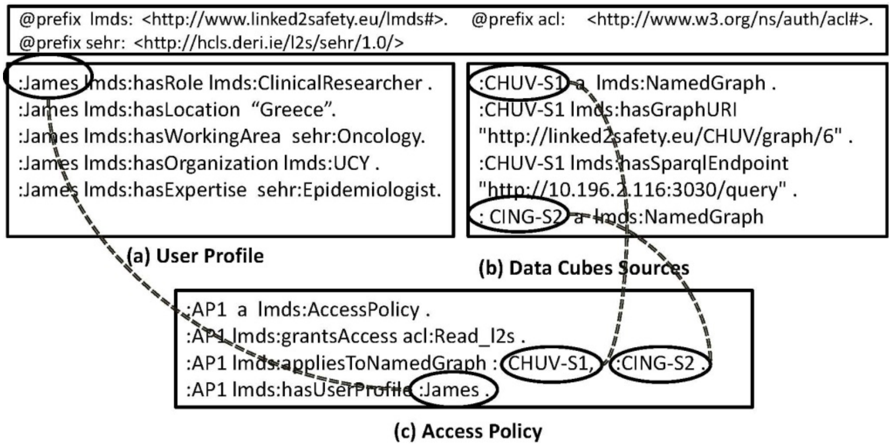


```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX sehr: <http://hcls.deri.ie/l2s/sehr/1.0/>
SELECT ?diabetes ?bmi ?hypertension ?cases
WHERE { ?dataset a qb:DataSet.
        ?observation qb:dataSet  ?dataset;
        a qb:Observation;  sehr:Diabetes ?diabetes ;
        sehr:BMI_Abnormal ?bmi ;
        sehr:Hypertension ?hypertension ; sehr:Cases ?cases . }
```

Figure 4 Example subject selection criteria for clinical trials

Referring back to Figure 3, only three of the datasets (*CHUV-S1*, *CING-S2* and *ZEINCRO-S3*) contain all required dimensions. An answer returned by the query (Figure 4) should list counts (*cases*) from these three RDF data cubes. However, assuming that the policy restrictions are applied to the user (say *James*), who wants to execute the query and has access to *CHUV-S1* and *CING-S2* RDF data cubes only. Therefore, the query federation engine should retrieve results only from *CHUV-S1* and *CING-S2* and should not consider *ZEINCRO-S3* for querying.

Hence, one of the key requirements in the context of the Linked2Safety project is to support federation of queries over clinical data distributed at multiple clinical sites by taking into account the data access policies (Figure 5 (c): shows a data access policy) assigned to the users (Figure 5 (a): shows a user profile for James) executing those queries. Since RDF data cubes are self-contained entities associated with additional provenance information (For example, creator, location, etc.; see Figure 5 (b)), they are modelled using named graphs [19] as supported in SPARQL: each named graph contains only one data cube and its provenance information.



```
@prefix lmds: <http://www.linked2safety.eu/lmds#>.   @prefix acl:   <http://www.w3.org/ns/auth/acl#>.
@prefix sehr: <http://hcls.deri.ie/l2s/sehr/1.0/>
```

```
:James lmds:hasRole lmds:ClinicalResearcher .
:James lmds:hasLocation "Greece".
:James lmds:hasWorkingArea sehr:Oncology.
:James lmds:hasOrganization lmds:UCY .
:James lmds:hasExpertise sehr:Epidemiologist.
```

```
:CHUV-S1 a lmds:NamedGraph .
:CHUV-S1 lmds:hasGraphURI
"http://linked2safety.eu/CHUV/graph/6" .
:CHUV-S1 lmds:hasSparqlEndpoint
"http://10.196.2.116:3030/query" .
: CING-S2 a lmds:NamedGraph
```

```
:AP1 a lmds:AccessPolicy .
:AP1 lmds:grantsAccess acl:Read_l2s .
:AP1 lmds:appliesToNamedGraph :CHUV-S1, :CING-S2 .
:AP1 lmds:hasUserProfile :James .
```

Figure 5 Example subject selection criteria for clinical trials

In order to publish clinical data cubes as RDF and describe user profiles along with their access rights used within query federation process, the Linked2Safety consortium has developed two vocabularies: (i) *Semantic EHR Model* (prefix "*sehr*") describes the clinical terminologies used by the three clinical partners; and (ii) *Access Policy Model* (prefix "*lmds*") describes the user profiles (their activity, location, organisation, position and role) and their respective access rights (for example, read, write). Considering space limitations, further details of these two vocabularies are out of scope for this paper; we instead refer the readers to dedicated papers on the Semantic EHR Model [16] and the Access Policy Model [18].

5.3. *The Platform*

The implementation of the integrated Linked2Safety platform concerned the in-depth design and development of all the modules of the generic reference architecture capitalizing upon the Service-oriented Architecture (SOA) paradigm. The main functionalities of the developed components are expressed in the form of web-services following Representational State Transfer (REST) architecture, and the communication among them is performed via serialized JavaScript Object Notation (JSON) messages which are encrypted.

The deployment of Linked2Safety involves the installation and configuration of a number of servers operating at a distributed scheme, with a clear separation of the infrastructure dedicated to the premises of the clinical data providers. This separation introduces the notions of the 'closed world room' and public access; on an attempt to further strengthening the security aspect of the infrastructure.

The concept of a 'closed world room' corresponds to the place located within a clinical data provider's premises, featuring the required hardware infrastructure to process EHRs isolated from any kind of network connections. The physical access to this machinery within the room is allowed only to specific personnel of the corresponding clinical data provider and it is off line to the outside world. The clinical data provider's personnel execute an application on the computers located in the closed room that aggregates the data generating the data-cubes. This application offers the option to the clinical data provider to limit the way that the data will be aggregated so that any legal and ethical issues that may relate to the type of analyses that may be performed on the data can be addressed so that the likelihood of reverse engineering of the data of a single subject or a group of subjects is eliminated. Quality control is also performed on the data by the application. Only the aggregated data (data-cubes) are physically carried outside the closed-world room computer to a server that is accessible by the rest of the Linked2Safety infrastructure. The data-cubes are in RDF format and can be accessed through a SPARQL endpoint, whereas all of the tools for analysing the aggregated data are on a dedicated Galaxy server for the needs of Linked2Safety. The analysis of the data is available only to Linked2Safety users through a Galaxy web portal. The source files of Linked2Safety platform, along with deployment and usage manuals are available to the public at https://github.com/linked2safety.

Moreover, in order to secure the communication protocols and message exchange within Linked2Safety a KPI infrastructure was employed. In specific, Layer-3 security protocols and VPN zones where introduced for securing the communication among the clinical data providers' premises hosting the data-cubes and the federated query engine (accessing those cubes), and between the clinical data providers and the Galaxy server (featuring the font-end interface for accessing Linked2Safety platform).

5.4. *The reference architecture*

The reference architecture designed following the data-cube concept defines a set of functional layers and the necessary components to support the operation of each layer for the realization of the platform that would enable the scalable, standardised, technical and semantic interconnection, sharing and reuse of heterogeneous EHR and EDC repositories in a secure way, facilitating the efficient homogenized access to and the effective, viable utilization of the medical information. Three of the most important layers spaces involve the:

Interoperable EHR Data Space which implements a toolset to make the transition of data-cubes from the 'closed world room' (generated locally within the premises) of each clinical data provider to an open data environment accessible to all users based on policies that enforce strong data security, privacy and anonymity. Thus, its responsibility is to transform the data-cube information to a common referenced data-cube format by means of a Semantic EHR Model, named Common data-cube Reference EHR Ontology. Moreover, the Interoperable EHR Data Space provides the mechanisms for the semantic enrichment of the standardized data-cubes with the use of appropriate, globally available healthcare and medical taxonomies and ontologies, enabling the delivery of machine interpretable information regarding their structure and content.

Linked Medical Data Space which realizes a secure Knowledge Base of semantically interconnected data-cube related information resources. It also provides the mechanisms and tools required for publishing and interlinking the common referenced data-cubes from different medical data providers, while links them with the Linked Data cloud. Access to this data is governed by adaptable access policies and mechanisms. In this way, the clinical research community has homogenized access to the available anonymized patient related information needed to perform complex data mining operations.

Data Analysis Space which provides a scalable infrastructure for medical data mining, empowered by a set of algorithms and models. These methods are applied in the semantically-interlinked data-cubes containing anonymized patient's health records in order to analyze the associations among the genetic, environmental and phenotypical data related to identified and reported AEs. Thus, clinical researchers and healthcare professionals are provided with an advanced genetic analysis statistical and data mining toolset focusing on advancing patients' safety through the analysis of bio-markers associated to identified AEs and the proactive exclusion of specific patients' profiles from the wide patients' selection process.

Additionally, the Conceptual artifacts layer plays an important role to the design of the reference architecture as it regards both the Common EHR Schema and Semantic EHR Model. The Common EHR Schema corresponds to a common-reference, interoperable EHR schema for aligning the open and widely adopted EHR standards and medical vocabularies such as the ISO/CEN 13606, the openEHR and the HL7-CDA (Clinical Document Architecture) needed for the semantically interlinking, sharing and reusing of clinical sources coming from distributed clinical data provider. This schema is used throughout the architecture for enabling the mapping of all proprietary and non-proprietary protocol-based EHRs and EDCs resources, and it is also utilized for the alignment of engines that facilitate the transformation of the EHR records and EDC databases to data to be used for creating genetic analyses.

The Semantic EHR Model forms another integral part of the architecture as it enables the seamless sharing among the authorized clinical data stakeholders participating in a clinical trial and linking of pieces of healthcare data i.e., EHR (clinical and healthcare data) and EDC (clinical trial system information). The model functions as a common ontological reference model for resolving ambiguity and heterogeneity of healthcare data (coming from distributed sources) used within the Linked2Safety environment. In addition, it is utilized in the Linked Medical Data Space for data-cube publication, facilitating the enrichment and annotation of heterogeneous data-cubes originating from different clinical data providers. It is important to note that the Linked2Safety Semantic EHR Model provided the foundation for the standardization of the Semantic Electronic Health Record (SEHR) ontology. The SEHR ontology is a light-weight and extensible ontology that covers multiple sub-domains of Healthcare and Life Sciences (HCLS) through specialization of the upper-level Basic Formal Ontology (BFO).

6. Evaluation Methodology

6.1. Internal evaluations

The first step was to conduct an internal evaluation, in which participants were clinical partners directly involved in the project. The internal user tests have been designed to validate that the platform as expected when used by experts from three different types of clinical research institutions, each supplying large, real-world medical datasets. They employ a variety of realistic use cases, fulfilling the use cases and key requirements of the system.

Internal users were members of the Linked2Safety consortium and had extensive knowledge of the aims, progress and delivery of the platform. This gave them significant in-depth knowledge. There were two groups of internal users: clinical partners and system developers, which are described next.

Clinical Partners (end-users) have been consulted extensively whilst developing the set of user requirements, and they have also designed the 'internal' showcases used to evaluate the Linked2Safety platform. The internal end-users have been fully involved in the scenarios, at their own place of work, performing tasks very similar to what they would expect to run when using the technology in real-world, non-test case scenarios, as they had full onsite access to the platform and data.

System developers were focused on the deployment, configuration, maintenance and further development of the Linked2Safety platform to enhance its usability and improve the service to the end-users. System developers carried out the deployment and configuration of Linked2Safety for new users or organizations and ensured that the Linked2Safety platform worked correctly and smoothly. Furthermore, they had provided technical support of the Linked2Safety platform to all partner users/organizations and have addressed technical issues with the Linked2Safety platform.

Cost deployment

Prior to conducting external evaluations, participants were provided data from the internal evaluation of Linked2Safety that involved real deployment and were asked to evaluating the economic impact associated with becoming a Linked2Safety Data Provider. Estimations for a prospective data provider include a total of 3 person months on preparatory activities, approximately

½ person month per 100 data variables for running activities and hardware costs which are estimated to be in the range of 500-1500 Euros.

More specifically, the preparatory activities include all those activities which a Data Provider will have to complete in order to participate in Linked2Safety. Preparatory expenses include the provision of two dedicated computers for the purposes of Linked2Safety; One for the closed room where the electronic patient data will be transformed into Data Cubes and one with network access to upload and access Data Cubes. In order for these preparatory activities to be efficiently and successfully executed, participation of a data manager, IT manager, legal advisor and project manager may be essential.

The running activities are all those activities which a company or organization will be carrying out once they have joined and are using Linked2Safety as Data Providers. Running activities are those activities which will be performed by most likely a research scientist once the organization has been affiliated with Linked2Safety as a Data Provider. These activities include the transformation of electronic patient data into Data Cubes (mapping of Data, production of RDFs) and the upload of Data Cubes onto the Linked2Safety Platform.

6.2. External Evaluations

The second step was to conduct external evaluations. External evaluators were members of the wider scientific community. External evaluations were designed and conducted to eliminate potential bias in the feedback and to gather a wider set of feedback on the utility and functionality of the Linked2Safety platform by people not affiliated with the project.

For the external user evaluations, a restricted second version of the integrated Linked2Safety platform was prepared and the installation of all components was done in a controlled environment. For security reasons, the external users did not have access to the full platform and could only analyze synthetic data [1] (modeled on real-world data). They were given specific scenarios to evaluate using demonstrations and screenshots, which tested the acceptance of confidence in the Linked2Safety concept and its applicability.

The idea behind scenario-based evaluation with these external users is for them to perform a type of task that is typical in their professional work by utilizing the Linked2Safety platform. Despite not having full access to the platform of the integrated real-world data, their insights were expected to have a non-biased view of the Linked2Safety platform, providing reliable and objective results.

6.2.1. Participants

The participants of the external evaluation came from three different groups: a) medical science analysts, b) analytic methodology engineers, and c) data providers, which coincide with the target groups of users of the system. These are described below.

Medical science analysts focus their efforts on analyzing data; they rely primarily on using existing statistical or computational methods to test pre-existing hypotheses or to generate new hypotheses depending on the problem on which they are working. They are typically associated with large pharmaceutical industry organizations, academic institutions interested in medical analyses and hospitals and other medical care providers that perform analyses on data as part of their decision support process, prognostics, or other efforts. Medical science analysts routinely seek new sources of data to test their hypotheses with increased statistical power, using standardized analytical tools.

Analytic methodology engineers are focused on developing innovative analytical techniques to perform analyses on data and on evaluating those techniques. They may have a background in statistics, computational intelligence, data mining, software engineering and development, or other fields of study. Their focus is on the development of tools that can either introduce new analytic approaches to solve medical problems through the analyses of medical data or to introduce new versions of analytic methodologies that are expected to have certain advantages over existing ones. Typically, an analytic methodology engineer would utilize the Linked2Safety platform as part of his/her efforts to evaluate newly developed tools and, once the tools are proven to be successful, the platform can also enable quick deployment of his/her work to a large number of medical science analysts for use.

Data providers are institutions that hold medical data; these may be organizations that are responsible for data collected through clinical trials, epidemiological studies, health providers with patients' electronic health records, and others who have the ability to store and use that medical data in some form of research analyses. The primary focus of these users is typically to collect data for scientific research, whilst strictly adhering to legal and ethical limitations.

Table V shows the three groups of participants of the external evaluations of Linked2Safety. There were a total of three external evaluation events with a total of 75 participants from the three targeted groups of potential users (35 medical science analysts, 11 methodology engineers and 29 data providers).

Table V External Evaluation Participants

Organising Partner	Count	Institution	Date	Types

[1] 'Synthetic data are often generated to represent the authentic data and allows a baseline to be set; another use of synthetic data is to protect privacy and confidentiality of authentic data.' (http://en.wikipedia.org/wiki/Synthetic_data)

CING	24	CING	May 2014	Medical science analysts
	23	CING	May 2014	Data Providers
CHUV	6	CHUV	July 2014	Medical science analysts
	6	CHUV	July 2014	Data Providers
UNIMAN	11	UNIMAN	July 2014	Analytic methodology engineers
	5	University of Liverpool	Aug 2014	Medical science analysts

Each evaluation event typically started with a brief description of the Linked2Safety project aims and scope, which was given as a presentation. The presentation included the results of the internal evaluation (e.g. monetary and time cost of deployment results, findings that replicated scientific knowledge already discovered) this was followed by individual hands-on experience of the Linked2Safety system by participants through three different scenarios (a different scenario for each group) that demonstrated the basic functionality of the system for each group of users' main activities. The workflow instructions were in the form of screenshots on how to use the basic functionality of the Linked2Safety platform. After participants used the system and had hands-on experience with its basic functionalities they completed an evaluation questionnaire (a different questionnaire for each group).

6.2.2. Questionnaire

The Linked2Safety external end-user evaluation questionnaire, which was developed specifically for the purposes of the evaluation of the Linked2Safety system, is structured in five main parts.

Part 1 of the evaluation questionnaire refers to users' personal information, such as gender, age, employment and experience.

Part 2 allows users to evaluate the following five aspects of the Linked2Safety platform:

A. Analyses space: Questions that fall under the category of analyses space cover issues of subject selection, hypothesis testing, hypothesis generation, data mining, replication testing, time, cost and usability.

B. Linked Data Space

C. Usability

D. Legal and ethical issues

E. Value of the system (for patients, future research)

The third part (Part 3) targets only members of Stakeholder Group 2 (analytic methodology engineers); and the fourth part (Part 4) targets only members of Stakeholder Group 3 (data providers).

Lastly, the fifth part (Part 5) of the external end-user evaluation questionnaire provides users with the ability to express their opinion in a few open-ended questions that focus on ways to improve Linked2Safety.

All questions were in a 'multiple choice' format and began with a statement to which the clinical partner was asked to state their agreement by selecting from the options: strongly disagree, disagree, agree and strongly agree. In addition, a 'not applicable' option was given should the user feel unable to give an answer. Not all statements were positive (a method commonly used in survey design to ensure that subjects do not attempt to reply completely positively or negatively without closely reading the questions).

6.2.3. Questionnaire Analysis

Following the data collection of the study, all the data from the different evaluation sessions were input in a statistical package (SPSS) for analysis. For the questionnaire results, descriptive statistics (frequencies) of 75 participants' answers have been collected to illustrate the users' perceptions of the Linked2Safety system, broken down into specific categories that can be analyzed in more detail (analyses space, linked data space, usability of the system, legal and ethical issues, value of the system).Associations/correlations between the users' personal data (e.g. their experience, educational level, age, gender) and their perceptions of the Linked2Safety system (e.g. to what extent they value the system, to what extent they find specific tools user-friendly etc.) were conducted. For these evaluations, a comparison of responses in questions that are common in the three target groups (data providers, medical science analysts and analytic methodology engineers) have been conducted to identify whether there are differences between users' perception of Linked2Safety based on their role.

7. RESULTS

Results in this paper focused on the external evaluation of Linked2Safety that represents an unbiased view of the potential scientific and societal impact of platforms such as Linked2Safety.

7.1 Participants Demographics

As summarized in Table V, there were three external evaluation events organized as part of the Linked2Safety project that took place in the following partners' premises:

a) CING (with 47 participants),
b) CHUV (12 participants), and
c) UNIMAN (16 participants).

The total number of external evaluators was 75, including 35 medical science analysts, 11 methodology engineers and 29 data providers. 45.3% of participants were male and 54.7% of participants were female. Over 90% of evaluators had either an MSc or PhD. The majority came from medical research institutes (42.7%) or academic institutions (30.7%). With regards to experience, almost half of the evaluators had at least 3 years of experience.

7.2 Results for five aspects of the Linked2Safety system

For all five aspects of the Linked2Safety system that were examined (analysis space, linked data space, usability, legal and ethical issues and value of the system), the participants' perceptions were overwhelmingly positive. These results are summarized in Table VI, which shows that all three groups had a Mean score between "agree" and "strongly agree" (higher than 3.10 out of 4.00 in all cases).

Table VI Descriptive Statistics of Participants' Perceptions in all Five Aspects of Linked2Safety

	N	Mean	Std. Deviation
Analysis space	75	3,2751	,49866
Linked data space	74	3,2195	,48834
Usability	75	3,0996	,63139
Legal and ethical	72	3,3264	,55131
Value of the system	75	3,2055	,43024
Valid N (listwise)	71		

Overall, the participants' perceptions of the analysis space were positive (Mean=3.28, SD=0.50) as all three groups had a Mean score between "agree" and "strongly agree".

The analysis space aspect includes all questions that refer to the overall functionality of the platform, as well as the functionality of the Linked2Safety system in relation to saving time and money when compared with traditional systems.

Considering the results in more detail, per group of participants, the descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 35) had the lowest scores in the evaluation of the Analysis Space (Mean = 3.22; SD = 0.45). Data Providers (N = 29) had the next highest evaluation score (Mean = 3.31; SD = 0.49) while Methodology Engineers (N = 11) had the highest evaluation score (Mean = 3.49; SD = 0.45). An ANOVA test showed no significant differences between the three groups (F = 1.423, p = 0.248) and thus no post hoc tests were run. There were also no significant differences among the groups in relation to gender, age, educational and employment.

A one sample t-test was then run to evaluate the inclination of answers (between satisfaction and dissatisfaction), with the value of 2.5 taken as the 'neutral' answer to be tested against for each group separately and all the groups together. The aim of this test is to examine whether the participants responses were in general higher than or lower than 2.5 (with 2.5 being the neutral answer between the values of 1 that showed disagreement/dissatisfaction with the particular aspect of the Linked2Safety system that was examined and 4 that showed agreement/satisfaction with the particular aspect of the Linked2Safety system that was examined), in other words it identifies whether the participants' responses were positive or negative at a statistically significant level. The results show a significant positive inclination ($p< 0.001$) in all groups separately as well as together.

At a more detailed level of analysis, we examined the descriptive statistics of individual questions that were part of the analysis space aspect. The vast majority of participants (over 90%) were positive about using the Linked2Safety platform to identify and combine data with other institutions and to locate datasets and subjects to test their hypotheses. Similarly, positive perceptions have been expressed about saving both money and time when deploying a new methodology (Net Per cent Agreement NPA=90.9%, n=75) or when locating data and selecting subjects (NPA=94.3%, n=75). Seventy per cent (70%) of evaluators agreed that the Linked2Safety platform could increase the statistical power of their experiments. While the participants were generally positive about using data mining to generate further hypotheses (NPA=91.9%, n=75), the scenarios they have executed, which used fake data, did not allow them to generate specific results that were worth investigating further, given the dataset used for external evaluation. They were also largely positive about using MedDRA for mapping of data (NPA=91.9%, n=75).

Furthermore, it was found that for the statement: 'I was able to investigate my hypothesis by testing for associations' only 5.7% disagreed and 94.3% agreed or strongly agreed. For the statement 'The use of Linked2Safety allowed me to successfully test the hypothesis of the study' 13.1% disagreed or strongly disagreed, and 84.7% agreed or strongly agreed.

7.1.1. Linked Data Space

Overall, the external evaluators' perceptions of the linked data space were positive (Mean=3.22, SD=0.49) as all three groups had a Mean score between "agree" and "strongly agree".

Considering the results in more detail, per group of participants, the descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 35) gave a medium evaluation score about Linked Data Space (Mean = 3.23) but with the highest variability in their answers (SD = 0.52). Data Providers (N = 29) gave the lowest evaluation score (Mean = 3.12; SD = 0.47) and Methodology Engineers (N = 10) showed the highest perception score (Mean = 3.33; SD = 0.44), which may be expected given their understanding and utilization of similar services for linking data. All three groups had a Mean score between "agree" and "strongly agree". An ANOVA test performed to investigate differences between the three groups showed no significant differences (F = 0.425, p = 0.656) and thus no post hoc tests were run.

A one sample t-test was run to see the inclination of answers (satisfaction and dissatisfaction), with the value of 2.5 taken as the 'neutral' answer: the results show a significant positive inclination ($p < 0.001$) in all groups separately as well as together. There were no significant differences among the groups in relation to gender, age, educational and employment.

At a more detailed level of analysis, we analysed the responses to individual questions that were part of the Linked Data Space aspect. The vast majority of answers (over 90%) reflect the perception that the linked data approach developed as part of Linked2Safety could provide a standardized and efficient way to enable merging of data from multiple sources for analysis. In addition, it could provide a meaningful and standardized approach to merging of clinical terminologies across multiple institutions (NPA=66.7%, n=75).

7.1.2. Usability

Overall, the external evaluators' perceptions of the usability of the Linked2Safety system were positive (Mean=3.10, SD=0.63) as all three groups had a Mean score between "agree" and "strongly agree". At a more detailed level of analysis, we analysed the responses to individual questions that were part of the system usability aspect. Overall, over 60% of evaluators thought that the platform was easy to use, and that the interface is not complex (NPA=90.6%, n=75). Around 80% of participants felt that the analytic space, the mapping tool, and the integration of MedDRA were easy to use, and similarly that the data mining tools were intuitive to use. The majority of evaluators expressed their motivation to use the Linked2Safety platform in the future (NPA=78.3%, n=75).

7.1.3. Legal and Ethical Issues

Overall, the external evaluators' perceptions of legal and ethical issues in relation to the Linked2Safety platform were positive (Mean=3.32, SD=0.55) as all three groups had a Mean score between "agree" and "strongly agree". The descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 33) gave the lowest evaluation score of legal and ethical issues (Mean = 3.24; SD = 0.59). Importantly, Data Providers (N = 29) gave the highest evaluation score (Mean = 3.43; SD = 0.53), while Methodology Engineers (N = 10) had a medium evaluation score (Mean = 3.30; SD = 0.48). Overall, all three groups had a Mean score between "agree" and "strongly agree". An ANOVA test performed to investigate differences between the three groups showed no significant differences (F = 0.914, p = 0.406) and thus no post hoc tests were run.

A one sample t-test was run to see the inclination of answers (satisfaction and dissatisfaction), with the value of 2.5 taken as the 'neutral' answer: the results show a significant positive inclination ($p < 0.001$) in all groups separately as well as together. There were no significant differences among the groups in relation to gender, age, educational and employment.

At a more detailed level of analysis, we analysed the responses to individual questions that were part of the legal and ethical aspect. Over 90% of evaluators felt that the platform guarantees anonymity of data (through data cubes) (NPA=93.2%, n=75), and almost 70% thought that re-identification of individuals was improbable using reasonable financial and technical efforts (NPA=69%, n=75).

7.1.4. Value of the System (for patients, future research)

Participants of the external evaluation were aware of the cost deployment results of the internal evaluation of the platform through a presentation of the system that preceded the administration of external evaluation instruments. Overall, the external evaluators' perceptions of the value of the Linked2Safety platform for patients and future research were positive (Mean=3.21, SD=0.43) as all three groups had a Mean score between "agree" and "strongly agree". The descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 35) had the lowest evaluation score on the value of the system (Mean = 3.12; SD = 0.43). Encouragingly, Data Providers (N = 29) had the next highest evaluation score on the value of the system (Mean = 3.27; SD = 0.40) and Methodology Engineers (N = 11) had the highest evaluation score on the value of the system (Mean = 3.36; SD = 0.40). Overall, all three groups had a Mean score between "agree" and "strongly agree".

Overall, over 80% of participants felt that the system enabled analyses that maximize the positive effect to the patient (NPA=80.6%, n=75). More than 80% of participants valued the version with MedDRA incorporated into the system, while two-thirds thought that the infrastructure would help them examine new analytical techniques, in particular if a large number of organisations is involved. Overall, the vast majority of evaluators (over 85%) thought that the Linked2Safety platform would have an impact on future research, is easy to use (NPA=75.8%, n=75) and that they would recommend it to other users and data providers (NPA=83.8%, n=75), in particular if it is made publically accessible (NPA=83.8%, n=75).

7.1.5. Participants' Motivation to Become a Data Provider

Overall, their composite score in these four questions indicates that they had a positive perception towards the idea of becoming a data provider (Mean=3.38, SD=0.46).). It is important to note that these results were positive even though participants were aware that there is some cost associated with the decision to become a data provider, which involves the need to employ a data manager, IT manager, legal advisor and project manager for the preparatory activities and a research scientist for the running activities. The group of Data Providers was also asked four specific questions to examine their motivation to become data providers for the Linked2Safety system.

Over 85% of data providers felt that the platform would reduce the time needed for data sharing compared to other approaches and that this should significantly increase the number of samples available (N=85.2%, n=75). Around two-thirds of evaluators thought that the process of becoming a data provider is simple (NPA=72.4%, n=75) and cost-effective (NPA=62.9%, n=75).

8. Discussion

From the deployment of the system that included three independent institutions from different European countries we were able to determine the costs for each aspect of the system deployment, as well as evaluate the new scientific potential of the increase in statistical power. It is clear that although only one data provider had whole genome data, while another had candidate gene studies, the overlap in genetic biomarkers as well as phenotypes was significant and allowed for the joint analyses of the datasets with a significant increase in statistical power. This indicated that a wider deployment of Linked2Safety or similar future system will results in significant gains in statistical power, enhancing the discoverability of new knowledge from existing data. Furthermore studies tend to collect a wide array of data beyond their primary and secondary endpoints that may be used either to adjust for known effects, or to study potentially unexpected/unknown effects. By merging data across many studies it's possible to gain sufficient statistical power to discover knowledge related to otherwise rare observations including combinations of biomarkers. Linked2Safety, clearly demonstrates a potential for such a system to enable re-use and a significant increase in data collected as part of isolated studies.

A set of external evaluation scenarios were developed for users working in the medical / pharmaceutical industry who are external to the project, in order to reduce possible bias. These were similar to the internal scenarios but, for security reasons, these external users did not have access to the full platform and could only analyze synthetic data. They were therefore given more general, less complicated scenarios to evaluate using demonstrations and screenshots which tested the acceptance of, confidence in and applicability of the Linked2Safety concept.

In all aspects the participants' responses were positive at a statistically significant level, as shown in the one-sample t-tests that were run. The results show a significant positive inclination ($p< 0.001$) in all groups separately as well as together, a finding that indicates that the acceptance level is high, their confidence in their ability to deploy the system in the future is high and, lastly their perceptions for the applicability of the concept are positive. In all cases there was no differentiation in participants' responses in relation to the group of users to which they belonged (analysts, methodology engineers or data providers). Further, in all cases, there were no significant differences among the groups in relation to gender, age, educational and employment.

If we look at the results per aspect, with regard to the analysis space, the vast majority of participants (over 90%) were positive about using the Linked2Safety platform to identify and combine data with other institutions and to locate datasets and subjects to test their hypotheses. Similarly, positive perceptions have been expressed about saving both money and time when deploying a new methodology or when locating data and selecting subjects. 70% of evaluators agreed that the Linked2Safety platform could increase the statistical power of their experiments.

With regard to the linked data space, the vast majority of answers (over 90%) reflect the perception that the linked data approach developed as part of Linked2Safety could provide a standardized and efficient way to enable merging of data from multiple sources for analysis.

With regard to usability, over 60% of evaluators thought that the platform was easy to use, over 90% of users thought that the interface is not complex and around 80% of participants felt that the analytic space, the mapping tool and the integration of MedDRA were also easy to use.

With regard to the legal and ethical issues, over 90% of evaluators felt that the platform guarantees anonymity of data (through data cubes).

Lastly, with regard to the value of the system, over 80% of participants felt that the system enabled analyses that increase the positive effect to the patient; while two-thirds thought that the infrastructure would help them to examine new analytical techniques, in particular if a large number of organizations is involved. The vast majority of evaluators (over 85%) thought that the Linked2Safety platform would impact future research and that they would recommend it to other users and data providers, as the system is currently publically available. It is important to note that participants' perceptions were positive while at the same time they were aware of the cost for deployment results of the platform.

9. Conclusion

This paper presented an innovative and secure semantic interoperability framework that is valuable for pharmaceutical companies, healthcare professionals and patients. Linked2Safety addressed the problem of diversity and complexity of today's legal and ethical regulations imposed at both the national and European legislation level, which make it difficult, risky and expensive to transfer data by sharing EHRs. The proposed solution provided a semantically interconnected approach to sharing aggregate data in the form of data cubes, which eliminated the risks associated with sharing pseudoanonymized (and therefore still personal in some types of data such as genetics) data while enabling the multi-source, multi-type analysis of health data through a single web based secure access platform.

The external evaluation that was conducted put the Linked2Safety theory into practice, allowing both clinical partners and potential external users coming from academia, and the medical and pharmaceutical industry to interact with the system. The research focus of the study was on the documentation of the perceptions of Medical science analysts, Analytic methodology engineers and Data providers on the evaluation of the system with respect to five specific dimensions (analysis space, linked data space, usability of the system, legal and ethical issues, and value of the system). For all five dimensions of the Linked2Safety system that were examined, the participants' perceptions were overwhelmingly positive, providing evidence of the acceptance of, confidence in and applicability of the Linked2Safety concept.

Linked2Safety or systems developed in the future based on similar concepts of aggregating data from multiple providers across Europe and beyond in a way that only research focus epidemiological analyses is enabled that adheres to national and international legal and ethical requirements could revolutionize the capacity for knowledge discovery without the need for larger, or significantly costlier studies. The tools exist already to enable standardization of data collected, as well as secure joint analyses; challenges remain however on the political and legal front with ambiguous and clustered legal frameworks. Consent seems to be of vital importance, as there is a lack of standardization on how the use of subject's data is limited to specific application (if at all) making efforts to enable wide ranging aggregate analyses challenging.

REFERENCES

[1] Catalina Martínez-Costa, Dipak Kalra, Stefan Schulz: Improving EHR Semantic Interoperability: Future Vision and Challenges. MIE 2014: 589-593

[2] Kalra, D; Schmidt, A; Potts, HWW; Dupont, D; Sundgren, M; De Moor, G; EHR4CR Research Consortium,; (2011) Case report from the EHR4CR project—A European survey on electronic health records systems for clinical research. iHealth Connections , 1 (2) 108 - 113.

[3] Amina Chniti, Lamine Traore, Sajjad Hussain, Nicolas Griffon, Stéfan Jacques Darmoni, Jean Charlet, Eric Sadou, David Ouagne, Eric Lepage, Christel Daniel: A Semantic Interoperability Framework for Facilitating Cross-Hospital Exchanges. MIE 2014: 1255

[4] Gokce B. Laleci, Mustafa Yuksel, Asuman Dogac, Providing Semantic Interoperability between Clinical Care and Clinical Research Domains, IEEE Transactions on Information Technology in BioMedicine, Volume: 17, Issue: 2, March 2013 (online since Sept. 2012), Page(s): 356-369.

[5] Gokce B. Laleci Erturkmen, Asuman Dogac, Mustafa Yuksel, Sajjad Hussain, Gunnar Declerck, Christel Daniel, Hong Sun, Kristof Depraetere, Dirk Colaert, Jos Devlies, Tobias Krahn, Bharat Thakrar, Gerard Freriks, Tomas Bergvall, Ali Anil Sinaci, Building the Semantic Interoperability Architecture Enabling Sustainable Proactive Post Market Safety Studies, Accepted as a poster in SIMI 2012 Wokshop (Semantic Interoperability in Medical Informatics), in ESCW 2012: Extended Semantic Web Conference, May 27, 2012 in Heraklion (Crete), Greece (Poster).

[6] R. Sahay, W. Akhtar, and R. Fox, "PPEPR: Plug and Play Electronic Patient Records," in Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil, March 16-20, 2008, R. L. Wainwright and H. Haddad, Eds. ACM Press, Mar. 2008, pp. 2298-2304.

[7] E. Directive, "95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," Official J. of the European Communities, vol. 281, pp. 31–50, 1995.

[8] R. Faden, T. Beauchamp, and N. King, A history and theory of informed consent. Oxford University Press, USA, 1986.

[9] "Privireal: Data protection - greece," Aug 2012. [Online]. Available:http://www.privireal.org/content/dp/greece.php

[10] "Office of the commissioner for personal data protection - home page," Aug 2012. [Online]. Available: http://www.dataprotection.gov.cy/dataprotection/dataprotection.nsf/d1813 d5911e138bdc2256cbd00313d1c/f8e24ef90a27f34f c2256eb4002854e7

[11] "Federal act on data protection," Aug 2012. [Online]. Available:http://www.vud.ch/generaldocs/vud revdsg/235.1 FADP en.pdf

[12] M. Firmann *et al.*, "The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome," *BMC Cardiovasc Disord*, vol. 8, p. 6, Mar. 2008.

[13] M. Preisig *et al.*, “The PsyCoLaus study: methodology and characteristics of the sample of a population-based survey on psychiatric disorders and their association with genetic and cardiovascular risk factors,” *BMC Psychiatry*, vol. 9, p. 9, 2009.

[14] A. Antoniades *et al.*, “The effects of applying cell-suppression and perturbation to aggregated genetic data,” in *IEEE 12th International Conference on Bioinformatics and Bioengineering (BIBE)*, Larnaka, Cyprus, 2012, pp. 644–649.

[15] Axel-Cyrille Ngonga Ngomo, Soren Auer. “LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data”. IJCAI, 2011

[16] R. Sahay, D. Ntalaperas, E. Kamateri, P. Hasapis, O. D. Beyan, M. F. Strippoli, C. Demetriou, T. Gklarou-Stavropoulou, M. Brochhausen, K. A. Tarabanis, T. Bouras, D. Tian, A. Aristodimou, A. Antoniades, C. Georgousopoulos, M. Hauswirth, and S. Decker. “An ontology for clinical trial data integration. In SMC”, pages 3244–3250. IEEE, 2013.

[17] Khan, Y., Saleem, M., Iqbal, A., Mehdi, M., Hogan, A., Hasapis, P. & Sahay, R. “SAFE: Policy Aware SPARQL Query Federation Over RDF Data Cubes”.

[18] E. Kamateri, E. Kalampokis, E. Tambouris, and K. Tarabanis. “The linked medical data access control framework”. Journal of Biomedical Informatics, 2014.