IMI1 Final Project Report Public Summary

Project Acronym: eTOX

Project Title: Integrating bioinformatics and chemoinformatics approaches for the development of expert systems allowing the in silico prediction of toxicities

Grant Agreement: 115002
Project Duration: 01/01/2010 - 31/12/2016

Executive summary

1.1 Project rationale and overall objectives of the project

Drug development necessitates the performance of *in vivo* toxicological studies for the identification of target organ toxicities, the assessment of potential side effects and the determination of a safe starting dose in first human trials. Toxicities may often limit the use of medicines, and sometime prevent molecules becoming drugs. Early selection of chemicals with a low probability of being toxic will improve the drug development process, taking less time and resources, including the use of animals. Hence, datamining in historical data and early *in silico* prediction of *in vivo* toxicological results would increase the efficiency of the drug development process and reduce the number of animals to be used in preclinical studies.

The eTOX project aimed to build a toxicology database relevant to pharmaceutical development and to elaborate innovative methodological strategies and novel software tools to better predict the toxicological profiles of new chemical entities in early stages of the drug development pipeline based on existing *in vivo* study results. This was achieved by sharing and jointly exploiting legacy reports of toxicological studies from participating pharmaceutical companies. The project coordinated the efforts of specialists from industry and academia in the wide scope of disciplines that were required for reliable modelling of the complex relationships existing between molecular and *in vitro* information, and the *in vivo* toxicity outcomes of drugs.

1.2 Overall deliverables of the project

The proposed strategy included a synergistic integration of innovative approaches in the following areas:

- Data sharing of previously inaccessible high quality data from toxicity legacy reports of the participating pharmaceutical companies;
- Database building, management and use, including procedures and tools for protecting sensitive data;
- Controlled terminologies (ontologies and code lists) and text mining techniques, with the purpose of facilitating knowledge extraction from and efficient usage of legacy preclinical reports and biomedical literature;
- Physico-chemical and structure-based approaches for the molecular description of the studied compounds, as well as of their interactions with the anti-targets responsible for secondary pharmacology;
- Prediction of DMPK (Drug Metabolism and Pharmacokinetics) features, often related to the toxicological events;
- Sophisticated statistical modelling tools required to derive relevant QSAR models.
- Development and validation (according to the OECD principles) of expert systems for the prediction of *in vivo* toxicity outcomes of drugs;
- Bioinformatics and systems biology approaches in order to cope with the complex biological mechanisms that govern *in vivo* toxicological events;
- Linkage of toxicity preclinical data with human outcomes.

All of the eTOX planned deliverables relate to these different types of activities that the project entails.

See the RESULTS site (http://www.etoxproject.eu/results.html) of the official project website for a compilation of freely available tools developed in the framework of eTOX as well as compiled dissemination activities.

1.3 Summary of progress versus plan since last period

The 7th reporting period has seen the consolidation of the preceding years of eTOX, and signifies the final leap forward for the project regarding all of its main objectives. The eTOX database has continued to grow, with 8,047 studies corresponding to over 7,000 legacy reports already extracted at the end of the project period.

Continuous attention to the quality of the data was devoted with revisions of the QC-tool developed, which detects and highlights possible errors and inconsistencies in the data which might occur during the data extraction. The use of this tool by companies and CROs in charge of data extraction showed up some other inconsistencies previously not identified and regular revisions helped to ensure a higher quality of the dataset.

The eTOX database schema was extended to allow capture of several new data types: Carcinogenicity (68 studies extracted), Safety pharmacology (43 studies extracted), Local Lymph-Node Assay (LLNA) (38 studies extracted). Further data gathering for genotoxicity and reproductive/developmental studies was deprioritized in order to run the quality checks

The eTOX ontology core team devoted dedicated efforts to include the latest SEND and INHAND controlled terminologies releases to improve the interoperability of eTOX data with standards guidelines. Following an SAB request, an extensive curation of the Pharmacological Action related information shared in some of the studies contributed was completed and glossary is also been used for the iPiEsys developed in the framework of the iPiE project[3].

The organisation of the second eTOX Hackathon structured in 3 'challenges', was again a successful transdisciplinary event in which a team of highly motivated members of the consortium, covering a wide expertise range (modellers, data experts, toxicologists and pathologists), collaborated closely together with the aim of developing models predicting relevant *in vivo* endpoints. Highly valuable results were produced during the Hackathon, which could have not been obtained otherwise.

The system developed for data browsing/searching and integration of the predictive models (eTOXsys) was optimized based on end user's active feedback during the last period to improve the potential use and retrieval of results (online and EFPIA in-house versions), incorporating several new features. The final release was launched online by mid December 2016, as a patch of version 3.0 released in mid-June 2016. The virtual machine version for installation behind the firewalls of the EFPIA partners was deployed in Jan 2017.

The prototype of the Human Outcomes module was improved in several aspects and a final version has been deployed to the consortium and fully integrated in eTOXsys.

All of the above has made eTOX become an increasingly rich, unique resource not only for early drug development but also interesting for synergising with other relevant initiatives worldwide. A specific methodology has also been adopted to identify, analyse and prioritise such projects, and facilitate implementation of mutually beneficial collaborations (OpenPHACTS and iPiE, both IMI initiatives).

To finalize the project, the main pillars (data and quality of the data, system improvement, and predictive models) were evaluated in terms of feasibility and usability. The consortium paid higher attention to the sustainability aspects covered by the Business Plan, key for the exploitation phase of the eTOXsys to the wide target audience in the drug development field. A full Business Plan is in place led by the business broker that will initiate its work from early 2017.

[3] Intelligence-led Assessment of Pharmaceuticals in the Environment (iPiE), http:// http://i-pie.org/

1.4 Significant achievements since last report

The 7th year of the project has seen eTOX consolidate its efforts with significant advances in the work performed for the tasks corresponding to ENSO and increasing normalisation of workflows affecting the pillars of the project (database, system, predictive models). Many of the activities have been aimed at optimising work around these pillars. The major achievements in the reporting period that can be highlighted are as follows:

- Version 3.0 of eTOXsys was released to the consortium, with the online and in-house versions fully operative. The eTOXsys application allows the querying of the eTOX database through standard chemical and text-based searches (both for toxicological effects and pharmacological mode of action) and allows installation of the system behind the firewalls of the EFPIA companies.

- The unique toxicology information database developed by the project has continued to grow, collecting legacy reports from industry partners and public data to develop better *in silico* tools for toxicology prediction of new compounds. At the end of the project, the pipeline includes over 7,500 preclinical reports already cleared by the participating pharma companies, with further reports being processed continuously – 7,000 having been extracted already. In addition, the data compiled from public sources covers about 265,500 substances (ChEMBL).

- Development of a QC-tool, which automatically detects and highlights potential errors and inconsistencies which can occur during data extraction, and reduces significantly the number of errors. Both companies and CROs made use and took advantage of this development for improvement of the data quality.

- Progress on model development, and, significantly, in methodologies enabling use of data and implementation of such models into the system, especially after the organization of the hackathons. Extended use of the eTOXlab modelling framework by most modelling partners for implementing their models. An early access to models procedure was created to overcome the time required for models to be deployed in eTOXsys. The outcome of modelling related activity has been a large collection of predictive models, along with a series of novel, reusable tools that support the use of predictive models in industrial environments.

- Development of the common ontologies to allow mapping of terms used across the different companies and also in public literature to a single preferred term, which is essential for cross-study data analyses, has also made very significant progress, with over 20 million entry lines. The OntoBrowser tool launched as open source to the community in early 2015 has been enriched already with recent publications of INHAND controlled terminology, regarding histopathology findings; and updated to SEND codelists recently released. Additional efforts devoted to ensure harmonization of the Pharmacological Action related information by mapping to preferred terms, aligned mostly with ChEMBL glossaries.

- Final version of the Human Outcomes Module was released and integrated in eTOXsys.

- A fully fledged eTOX Business Plan was consolidated and a User Board was formed, starting its operations in the market from early 2017.

1.5 Scientific and technical results/foregrounds of the project

ID	Title	Description	Owner	Licensing / expl. information
F01	ChOX database	ChOX is a database containing public-domain bioactivity data of relevance to the eTOX project, primarily extracted from the scientific literature. The database includes 2-D structural and physicochemical information on molecules as well as information about their bioactivities including their interaction with molecular targets and their pharmacokinetic and toxicological properties. Scripts have been created to run automatically extraction from the ChEMBL database following each release. The integration into eTOX database allows for a consistent access to data in a single interface	EMBL	ChOX is loaded directly in Vitic through an SDfile formatting data from ChEMBL (Lhasa could manage the distributing/hosting of the ChOX database). All the ChOX data is available under the same licence terms as ChEMBL, Creative Commons Attribution-Share Alike 3.0 Unported License (https://www.ebi.ac.uk/chembl/about)
F02	eTOX database / Vitic	Chemical centered database compiling all data gathered in the scope of the project	Lhasa Limited	Licensing of the non-confidential part of the eTOX database to third parties (e.g. allocation of royalties). It is assumed that the data and the database structure will be part of a single offering. *Note of clarification: the data remains the companies' background/ownership.*
F03	eTOXsys	eTOX integrated system	MN	"License pack" with other eTOXsys components (potentially the database, the server, the modules, background and third party licenses)
F04	Predictive modules	Each predictive module is a blind prediction engine producing a predicted assessment for a given toxicity endpoint	Modeller partners	
F05	Ontologies	Ontologies are formal representation of knowledge within a specific domain that show the relationships between different concepts in that domain. The eTOX ontologies compile verbatim terms extracted from the nonclinical study data reports according to specific	Novartis	Will be made available open source.

ID	Title	Description	Owner	Licensing / expl. information
		fields relevant for the project. SEND and INHAND controlled terminologies are used as standards.		
F06	OntoBrowser	Tool developed to manage ontologies and controlled vocabulary used within the scope of eTOX. The primary functionality is to provide an online collaborative solution for expert curators to map the verbatim study data report terms (from the eTOX database) to preferred ontology (or controlled terminology) terms.	Novartis	Available open source under Apache License, version 2.0 (http://opensource.nibr.com/projects/ontobrowser/)
F07	eTOXlab	Flexible modelling framework, developed for supporting models predicting the biological properties of chemical compounds in production environments	FIMIM	Available open source under license GNU GPL version 3 (http://phi.imim.es/envoy/)
F08	eTOXvault	Database containing predictive models documentation	FIMIM	Integrated in eTOXsys
F09	Structure Standardizer	Tool designed to provide a simple way of standardizing molecules as a prelude to e.g. molecular modelling exercises.	EMBL	Available open source released under the Apache 2.0 license (https://github.com/flatkinson/standardiser)
F10	LimTox	Text mining approach that extracts associations between compounds and toxicological end points (hepatotoxicity, nephrotoxicity, cardiotoxicity, thyroid toxicity and phospholipidosis) at various levels of granularity and evidence types, all inspired by the content of toxicology reports.	CNIO	Available open source (http://limtox.bioinfo.cnio.es/)
F11	Collector	Tool developed in collaboration with the IMI OpenPhacts project, which allows extraction from the OpenPhacts Platform series of biological annotated compounds that can be used directly for building predictive models	FIMIM	Implemented in eTOXlab under GPL v3 license (http://phi.imim.es/collector/)
F12	ADAN (Applicability	Method for the assessment of the prediction reliability based on an exhaustive description of the	FIMIM (UNIVIE)	Implemented in eTOXlab or as a separate R package and distributed under GPL v3 license

ID	Title	Description	Owner	Licensing / expl. information
	Domain Analysis)	model applicability domain (article: http://www.ncbi.nlm.nih.gov/pubmed/24821140)		
F13	Tox-PPI	Network visualization tool to facilitate the navigation through the protein-protein interactions (PPI) networks expanded with pharmacological and toxicological data available in the public domain, aiming at exploring the molecular mechanisms of drug toxicity	DTU (FIMIM)	Available open source (https://www.cbs.dtu.dk/projects/Tox-PPI)
F14	Database of disease-related biomarkers	A knowledge-driven text mining approach that can exploit a large literature database to extract a dataset of biomarkers related to diseases covering all therapeutic areas	FIMIM	Available open source under the Open Database License. (http://ibi.imim.es/biomarkers/)
F15	eTOX Library	Compilation of selected Articles and Journals extracted from ISI Web of Knowledge with Toxicology filtering; and Links to toxicology related web resources	FIMIM	Available open source (http://cadd.imim.es/etox-library/)
F16	Verification Models workflow	The workflow comprises an assessment of (i) the quality of data used to build (and test) the model, (ii) whether the predictions generated by the model when executed in eTOXsys are consistent, and (iii) the completeness of the documentation accompanying the model. This process has been semi-automated by implementing the required checks into a KNIME workflow and results in a more efficient and standardized verification process. The workflow requires the training data comprising structure ID, SMILES, experimental activity and predicted activity, along with specified outputs from eTOXsys. The result of the workflow is a PDF report that includes the verification status (Verified or Not Verified) and details on all of the verifications checks which were performed.	LMJU	Available open source (https://drive.google.com/file/d/0B70msg9CQ8R1MTlXN3dEb0ppZ0E/view)

ID	Title	Description	Owner	Licensing / expl. information
F17	Human Outcomes Module	A webservice that allows the search of similar marketed drugs from a simple SMILES code and retrieves adverse events data associated to these drugs from three sources currently implemented: literature (PubMed), EURETOS and LAERTES.	ERASMUS FIMIM	Available open source (http://euadr.erasmusmc.nl/etox2client/)
F18	eTOXsys sampler	To assist in marketing eTOXsys a sampler database is also planned to be released containing 58 substances and 96 studies donated by Bayer, BI, GSK, Janssen, Novartis, Roche, SAD & UCB (https://etoxsys.eu). The data was anonymised and clinical signs data excluded.	MN	

1.6 Potential impact and main dissemination activities and exploitation of results

Over the seven years of duration, eTOX has generated several results, some of them being publicly available through the eTOX website (http://www.e-tox.net/results.html).
eTOX members have been very proactive in disseminating the project; the almost 400 dissemination activities are good evidence of this. Some of these activities can be highlighted, as the 9 posters presented at the Society of Toxicology in 2016 or the symposium that will be organised at EuroTox 2017, with 5 speakers from both inside and outside the project, entitled *Beyond data sharing - towards data transparency, management, mining and application to predictive safety assessment* (http://www.eurotox2017.com/monday/session-6/).

During the second Hackathon held in March 2016, a video was prepared explaining the project and the event. It has obtained notable interest with almost 600 visits in youtube.com already.

eTOX has entailed a unique endeavour by which a consolidated standardized toxicity database consisting of preclinical data from both marketed drugs and failed drug candidates has been developed. The establishment of the data sharing procedures has meant a quantum leap forward in the way that pharmaceutical companies work together. In addition, the creation of a system which combines the data with the prediction models allows to the companies to implement new methodologies in how drug candidates are built.

The project has developed a Business Plan to guarantee the exploitation of the results. The business broker through a User Board will be in charge of implementing the sustainability phase of the project enabling its continuity during the exploitation phase. The User Board will determine the fees for future customers based on fair and reasonable terms, i.e. the access and maintenance fees will consider aspects such as new data contribution (in-kind contribution), the number of users per institution, the role of the institution (commercial vs academia/regulatory) and, if commercial, the size of the company. The commercialization of the database is intended immediately after the end of the project, i.e. during the first quarter of 2017. The initial presentation of the system outside the project will be done during the Annual Conference of the Society of Toxicology held in March 2017.

1.7 Lessons learned and further opportunities for research

In its seven years of duration, eTOX has developed a paramount experience and also learned from the difficulties encountered in all the period. As stemming from Call 1 in IMI1 and being the first project in Translational Safety, eTOX has set the scene in several aspects.

- Honest Broker

The Honest Broker concept was established in order to guarantee the management and secure access to the data provided by the EFPIA companies. Bi-lateral non-disclosure agreements between the Honest Broker and the companies were established. This figure has been also used in projects initiated after the foundation of the eTOX project.

- Data sharing

The classification of the data between confidential and non-confidential was an issue left at the discretion of each company that meant a delay to data access from model developer's side. In future initiatives, a reasonable target for the rate of shared (non-confidential) data for read-across or modelling purposes should be established from the early beginning. In addition, confidential

agreements should be taken as an advantage to minimize the effect of pending legal clearance to undergo activities.

- Data integration

An evaluation of a conventional and already existing type of data cleared in the framework of the project was used for a first database schema. Based on a list of possibilities on how to query the database, the database schema was subsequently modified several times with the perspective to support more efficiently the decisions in the business drug development pipeline. New contributions among the project period implied additional revisions of that first schema, mainly intended to allocate new types of data or diversity in data inputs.

In order to assure the optimal exploitation of any integrative system, like the eTOX database or eTOXsys, there was a need to establish a curation procedure looking for high interoperability between sources combined to offer a harmonized retrieval of results for a more simple and efficient evaluation of results by end users. For instance, the Controlled Terminologies Management involving experts from different disciplines, to map partner terminologies to a common preferred terms list, was key to make the database compatible with standards (i.e. SEND code lists, INHAND publications) and to notably improve the data quality for read-across analysis and modelling purposes.

The quality issues encountered during data extraction and data use were collected and rules for Best Practice on data processing were reported. To correct the inconsistencies, QC tools were developed to cover the issues identified. An earlier exhaustive checking of data quality is advised from the beginning for two main reasons, to have a full evaluation of data types to define the database schema and to ensure data reliability for read across analysis and modelling purposes.

- Data access strategy

Aiming at facilitating data sharing with external collaborators (i.e., other ongoing IMI projects, consortia applying to a Horizon 2020 call, US initiatives (ToxCast, Tox21), single organisations or funding agencies), a specific eTOX data access strategy was developed to provide a framework to deal with requests in a consistent manner (Cases et al. Int. J. Mol. Sci. 2014, 15).

In fact, such strategy should have been implemented from the beginning of the project also for requesting confidential data by model developing partners and speed up accessibility, since the clearance of such amount of data blocked the earlier use of data.

- Workshops and Hackathons

In order to enable communication and joint efforts between modellers and toxicologists, several workshops and hackathons were organised, which facilitated easier and faster understanding of both sides capabilities, interests, needs and limitations. Grouping different experts to solve a common challenge served as a pilot exercise or starting point to define project task forces and discuss optimal actions to reach the project goals taking benefit of the consortium public private partnership. This type of meetings proved to be effective collaboration points between the different working groups in the project.

Additionally, the workshops (2-4h sessions as satellite meetings to Consortium meetings) held were perceived as being highly valuable to discuss modelling methodologies, identify technical aspects to cover the models development, or to define common resources to support the integration of the models in the eTOXsys platform.

In the case of the Hackathons, the format of 4-days meeting was sufficient to reach a common understanding of the challenge proposed, explore options to overcome it in a cooperative way and process and analyse data to make progress on models building. As an example, during the first hackathon, the fact of having an open challenge with a multidisciplinary team was conclusive to identify subsets of data interesting for specific use cases that needed further curation.

- Sustainability concept

Sustainability aspects were not addressed in the initial proposal of the project, and were only raised as part of the common goals of the project with the ENSO extension.
Based on the ongoing and expected outcomes from the project, a Sustainability Business Plan was developed by key partners of the consortium (honest broker, eTOXsys joint developer, project managers and ExCom board). To turn first thoughts into facts, several meetings were held and dialogue with the rest of consortium members was undertaken build a mature plan. Market analysis and a list of feasible outcomes of the project proved the challenges of the eTOX business plan.

After a market analysis, a modular approach was pointed out as the optimal business like solution for stepping in in the field of the decision support software, as a cooperative integration of resources to engage interest from a wider audience, not only research community but also regulatory bodies.

Strategic sustainability steps must be carefully considered to be planned earlier in the project timeline, combining both short and long-term strategies and moving forward in parallel to the project activities. Decisions on project outcomes development might be influenced by sustainability aspects (i.e. in the case of model building, access to third party software might be critical for example in model maintenance concerns).

- General aspects

In general, a good legal coordination in all project phases is a strategic recommendation for any initiative involving sharing and re-use of archived data. For example, some EFPIA partners felt short of contributions, or additional efforts were needed to define procedures for setting user requirements, system specification and verification. A nomination of a Chief Technical Officer (CTO) might have helped in the coordination of the project precisely to speed up how to overcome the diverse bottlenecks in some of the key issues in the timeline of the project.