Legacy data sharing to improve drug safety assessment: the eTOX project

Ferran Sanz, François Pognan, Thomas Steger-Hartmann, Carlos Díaz and eTOX*

**See online for full author list.*

Full list of authors for the online version:

Ferran Sanz, François Pognan, Thomas Steger-Hartmann, Carlos Díaz, Montserrat Cases, Manuel Pastor, Philippe Marc, Joerg Wichard, Katharine Briggs, David Watson, Thomas Kleinöder, Chihae Yang, Alexander Amberg, Maria Beaumont, Anthony J. Brookes, Søren Brunak, Mark T. D. Cronin, Gerhard F. Ecker, Sylvia Escher, Nigel Greene, Antonio Guzmán, Anne Hersey, Pascale Jacques, Lieve Lammens, Jordi Mestres, Wolfgang Muster, Helle Northeved, Marc Pinches, Javier Saiz, Nicolas Sajot, Alfonso Valencia, Johan van der Lei, Nico P.E Vermeulen, Esther Vock, Gerhard Wolber and Ismael Zamora.

Affiliations:

Ferran Sanz, Manuel Pastor
Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Universitat Pompeu Fabra, 08003 Barcelona, Spain

François Pognan, Philippe Marc
Novartis Institute for Biomedical Research, Basel, CH-4002, Switzerland

Thomas Steger-Hartmann, Joerg Wichard
Bayer AG, 13353 Berlin, Germany

Carlos Díaz, Montserrat Cases
Synapse Research Management Partners, 08007 Barcelona, Spain

Katharine Briggs, David K. Watson
Lhasa Ltd, LS11 5PS Leeds, United Kingdom

Thomas Kleinöder, Chihae Yang
Molecular Networks GmbH, 90411 Nürnberg, Germany

Alexander Amberg
Sanofi, 65926 Frankfurt am Main, Germany

Maria Beaumont
GlaxoSmithKline Research and Development Ltd

Anthony J. Brookes
University of Leicester, LE1 7RH Leicester, United Kingdom

Søren Brunak
Technical University of Denmark (DTU), 2800 Lyngby, Denmark

Mark T. D. Cronin
Liverpool John Moores University, Liverpool L3 3AF, United Kingdom

Gerhard F. Ecker
Universität Wien, 1090 Vienna, Austria

Sylvia Escher
Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM), 30625
Hannover, Germany

Nigel Greene
Pfizer Ltd, Groton, CT 06340, USA
(Current affiliation: AstraZeneca, Waltham, MA 02451, USA)

Antonio Guzmán
ESTEVE, 08041 Barcelona, Spain

Anne Hersey
European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome
Genome Campus, Hinxton, Cambridge, CB10 1SD United Kingdom

Pascale Jacques
UCB Pharma SA, 1070 Brussels, Belgium

Lieve Lammens
Janssen Pharmaceutica NV, 2340 Beerse, Belgium

Jordi Mestres
Chemotargets SL, 08003 Barcelona, Spain

Wolfgang Muster
F. Hoffmann-La Roche AG, CH-4070 Basel, Switzerland

Helle Northeved
H. Lundbeck A/S, 2500 Copenhagen, Denmark

Marc Pinches
AstraZeneca AB, SK10 2NA Cheshire, United Kingdom
(Current affiliation: Lhasa Ltd, LS11 5PS Leeds, United Kingdom)

Javier Saiz
Universitat Politècnica de València, 46022 València, Spain

Nicolas Sajot
Servier, 45000 Orléans, France

Alfonso Valencia
Centro Nacional de Investigaciones Oncológicas, 28029 Madrid, Spain
(Current affiliation: Barcelona Supercomputing Center, 08034 Barcelona, Spain)

Johan van der Lei
Erasmus Universitair Medisch Centrum, 3015 CE Rotterdam, The Netherlands

Nico P.E Vermeulen
Vrije Universiteit Amsterdam, 1081 Amsterdam, The Netherlands

Esther Vock
Boehringer Ingelheim International GmbH, 88379 Biberach an der Riss, Germany

Gerhard Wolber
Inte:Ligand GmbH, 1070 Vienna, Austria

Ismael Zamora
Lead Molecular Design S.L, 08172 Sant Cugat del Vallès, Spain

Correspondence to:

F. S. & F. P.
ferran.sanz@upf.edu; francois.pognan@novartis.com

Preface:

The sharing of legacy preclinical safety data among pharmaceutical companies and its integration with other sources of information offers unprecedented opportunities for early safety assessment through read-across and predictive modeling. The experience of the IMI eTOX project is presented.

Subject terms: Drug Safety. Toxicology. Data Sharing. *In silico* prediction.

Non-clinical safety assessment is often faced with the challenge of assessing candidate compounds with no or insufficient experimental data. Whereas relevant databases and reliable *in silico* tools exist for mutagenicity prediction, analogous resources for identifying potential organ toxicities are not as common or well developed. Scientists, therefore, have to resort to suboptimal procedures such as literature search and personal experience with similar compounds or classes. In parallel, there is a wealth of relevant data buried in the archives of the pharmaceutical industry that has not yet been leveraged. These data mainly exist in paper or pdf formats and, consequently, are difficult to search and analyze. In order to overcome these limitations and advance early safety assessment, 13 pharmaceutical companies, 11 academic partners and 6 small and medium-sized enterprises (SMEs) joined forces in the eTOX project, which started in January 2010 under the sponsorship of the European Innovative Medicines Initiative (IMI). Since the availability of a wide and representative collection of historical data is fundamental for generating reliable predictions, the main goals of the eTOX project were (i) to build a shared and mineable database containing a broad and relevant collection of data, constituted mainly by repeat-dose toxicity studies contributed by the pharmaceutical companies participating in the project, and (ii) to use the database and other sources of information for enabling more effective read-across and predictive modeling of safety endpoints.

The eTOX database and other sources of information

At the end of the project (December 2016), the eTOX database contained 8,196 extracted toxicity studies carried out on 1,947 compounds, which report nearly nine million preclinical data points (in life observations, gross necropsy, histopathology, laboratory values like clinical chemistry, hematology, urinalysis). The database contains repeat-dose studies executed in different species (rat: 58%; dog: 28%; monkey: 7%; mouse: 5%), spanning different study durations (74% lasting up to 28 days; 17% between 29 and 91 days; 9% longer duration, up to 2 years) and diverse administration procedures (oral: 77%; intravenous: 14%). For a significant fraction of the compounds complementary information is included, such as pharmacokinetics (67%), pharmacological target (44%) and therapeutic indication (16%). The chemical structure is disclosed for most (75%) of the compounds present in the database. Nearly 20% of the eTOX compounds are FDA approved drugs and the chemical space covered by the eTOX database represents well the internal chemical spaces of the pharmaceutical companies as determined by PCA on above 100 molecular descriptors. This feature is essential for obtaining reliable predictions and performing meaningful read-across. Moreover, all pairwise Tanimoto coefficients were calculated to compare the structural similarities within eTOX and FDA drugs sets, and across the two sets. The structures of FDA drugs were, on average, significantly ($p<0.001$) more correlated with those in the eTOX set than among them. Another important feature for predictivetiness is the presence of adequate numbers of positive and negative compounds (e.g. considering four weeks sub-chronic rat studies, 965 compounds out of 1854 caused treatment-related effects in liver and 870 in kidney). The eTOX database is hosted by an *honest broker* tasked with providing adequately protected access to the data, enabling the sharing of sensitive information among pharmaceutical companies. A small subset of database is made available for public use from the eTOXsys Dashboard.

The eTOX database and other data resources, such as RepDose, ChEMBL and DrugBank, have been used by eTOX partners for developing nearly 200 predictive models (*in vivo* organ toxicity: 12 models; *in vitro* assays for organ toxicity: 30; safety pharmacology: 97; transporters: 26; ADME: 29; etc.), in which innovative modeling strategies have been applied (1,2). The extraction and integration of information from public sources for model development has been facilitated by an *ad hoc* tool (Collector), which operates on top of the IMI Open PHACTS data infrastructure. In addition, the extraction of relevant information from free text (scientific literature, EPAR reports) required the development and application of text mining techniques (3).

eTOXsys: Integrated platform for read-across and prediction

The eTOX database and predictive models can be accessed using an integrated and user-friendly software platform (eTOXsys). This platform incorporates a graphical user interface (GUI) that allows performing sophisticated queries by providing an easy-to-use query builder to interrogate different criteria including substructure, structural similarity, study characteristics, finding type, etc. The GUI also allows selecting and executing the available models, presenting the predictions in a way that facilitates their visualization and interpretation by providing model metadata. eTOXsys facilitates the export of queries and predictions in commonly used formats. Moreover, the system incorporates an exploratory module that queries possible human safety liabilities for a compound on the basis of its similarity to marketed drugs based on profiles of both structures and toxicity.

Use and impact in the pharmaceutical industry

eTOXsys has not only found its entry into the daily practice of early drug safety assessment but also in later phases, such as the assessment of impurities in drug products that have not been qualified for *in vivo* studies. The main application lies in the comparison of hit or lead structures with compounds in the database to raise hypotheses regarding organ toxicities (read-across). Given that the database can be searched for a particular pharmacological target, an eTOX pharmaceutical company observed that the same target has been previously addressed by other companies in more than 15% of the compounds stored in the database. This kind of information can help assess whether observed toxicities in early *in vivo* studies are target-related or off-target. In certain circumstances, the data can be used to make head-to-head comparisons with competitor compounds without performing additional animal studies, thus contributing to 3R (Refine, Reduce, Replace) policies.

Furthermore, the database can be used to analyze the correlation between the presence of chemical substructures and the occurrence of specific toxicities, e.g. if a chemical moiety raises the suspicion of being related to hepatotoxicity, the substructure can be queried in the database in combination with liver findings. The result of this search can be compared with the background occurrence of the liver findings for all compounds. If a significantly higher incidence is found for the substructure, then there is high probability of causality, which can be managed by the medicinal chemists by modifying the substructure without discarding the whole molecule.

Independently of the use in drug development, the eTOX database is also contributing to the refinement of thresholds of toxicological concerns (TTC). A recent analysis of the non-observed adverse effect levels of systemic toxicity studies in eTOX supports the ICH Q3A defined concept that a lifetime dose to 1 mg/day of a non-mutagenic impurity would not represent a safety concern for patients (3). This study also provided evidence that higher thresholds can be accepted for shorter administration periods during early clinical development.

Challenges, lessons and future perspectives

The first challenge to be overcome by the eTOX consortium was the apprehension of the pharmaceutical companies to share sensitive proprietary pre-clinical data. This required a combination of legal (consortium agreement), technical (e.g. database installed behind companies' firewalls and models implemented within self-contained virtual machines), organizational (honest broker concept), psychological (trust gained through collaboration), political (data sharing pressure, such as the FAIR principles) and social (snowball effect) solutions.

The second challenge was the lack of standardization in the data contributed by the pharmaceutical companies. This required the collaborative development and implementation of relevant ontologies. The project developed a specific software tool (Ontobrowser) for facilitating such cooperative task. Using this tool, 105,000 verbatim finding terms were mapped to 7,300 preferred terms aligned with existing CDISC and INHAND terminologies to enable interoperability (see Hpath in S2). We also carried out data curation campaigns to increase the quality of the data incorporated into the database. Moreover, the raw data extracted from the reports was not directly appropriate for number crunching and had to be transformed into comparable toxicity scores suitable for modeling. Quick progress was obtained in *hackathons* organized by the eTOX consortium, which defined rules for summarizing related toxicological findings and identified underlying relationships between chemical structures and organ toxicities (e.g. causes of bilirubinemia).

The development of a high number of models by diverse organizations, their integration into a single system (eTOXsys) and their adoption by the pharmaceutical industry, was another challenge that required strict procedures and tools for enabling a harmonized development, documentation, validation and implementation of the models, as well as their versioning and maintenance. These tasks required the development of new tools including eTOXlab, a flexible modeling framework, ADAN (4), a method for the assessment of the model applicability domain, and specific protocols for model validation (5). Models were extensively documented following OECD QSAR reporting recommendations and stored in the eTOXsys Model Meta Database. Executive summaries are accessible though the eTOXsys GUI for facilitating the interpretation of the predictions on the fly.

The eTOX IMI grant finished on December 2016 and the project entered into its sustainability phase with SME partners leading the commercial exploitation of eTOXsys. A User Board with representatives of the different partner oversees the maintenance and exploitation processes. Finally, it should be stressed that recently funded IMI projects are extending the *in silico* toxicology domain furthered by the eTOX project. This is the case of the TransQST and eTRANSAFE projects, which show a clear focus on animal to human translation. TransQST is devoted to translational quantitative systems toxicology, while the eTRANSAFE project will push forward guidelines for legacy data sharing, and will jointly exploit preclinical data (mostly in CDISC format) and clinical safety information for a better prediction of potential human safety liabilities.

References

1. Obiol-Pardo, C. *et al.* A Multiscale simulation system for the prediction of drug-induced cardiotoxicity. *J Chem Inf Model.* **51**, 483-492 (2011).
2. Carbonell, P. *et al.* Hepatotoxicity prediction by systems biology modeling of disturbed metabolic pathways using gene expression data. ALTEX. **34**, 219-234 (2017).
3. Harvey, J. *et al.* Management of organic impurities in small molecule medicinal products: Deriving safe limits for use in early development. *Regulat Toxicol Pharmacol.* **84**, 116-123 (2017).

4. Carrió, P. *et al.* Applicability Domain Analysis (ADAN): A Robust Method for Assessing the Reliability of Drug Property Predictions. *J Chem Inf Model.* **54**, 1500-1511 (2014).
5. Hewitt, M. *et al.* Ensuring Confidence in Predictions: A Scheme to Assess the Scientific Validity of In Silico Models. *Adv Drug Deliv Rev.* **86**, 101-111 (2015).

Additional references (S1) and web links (S2) provided as Supplementary Information

Competing interests statement

Some of the authors are employed in the pharmaceutical industry or in SMEs, as indicated in their affiliations.

Supplementary material

S1. Extended list of bibliographic references:

- Obiol-Pardo, C. *et al.* A Multiscale simulation system for the prediction of drug-induced cardiotoxicity. *J Chem Inf Model.* **51**, 483-492 (2011).
- Briggs, K. *et al.* Inroads to predict in vivo toxicology-an introduction to the eTOX Project. *Int J Mol Sci.* **13**, 3820-3846 (2012).
- Cases, M. *et al.* The eTOX Library of Public Resources for in Silico Toxicity Prediction. *Mol Inform.* **32**, 24-35 (2013).
- Martí-Solano, M. *et al*. Integrative knowledge management to enhance pharmaceutical R&D. *Nat Rev Drug Discov.* **13**, 239-240 (2014).
- Cases, M. *et al.* The eTOX data-sharing project to advance in silico drug-induced toxicity prediction. *Int J Mol Sci.* **15**, 21136-21154 (2014).
- Carrió, P. *et al.* Applicability Domain Analysis (ADAN): A Robust Method for Assessing the Reliability of Drug Property Predictions. *J Chem Inf Model.* **54**, 1500-1511 (2014).
- Hewitt, M. *et al.* Ensuring Confidence in Predictions: A Scheme to Assess the Scientific Validity of In Silico Models. *Adv Drug Deliv Rev.* **86**, 101-111 (2015).
- Carrió, P. *et al.* eTOXlab, an open source modeling framework for implementing predictive models in production environments. *J Cheminform.* **7**, 8 (2015).
- Sanz, F. *et al.* Integrative Modeling Strategies for Predicting Drug Toxicities at the eTOX Project. *Mol Inform.* **34**, 477-484 (2015).
- Garcia-Serna, R. *et al.* Large-Scale Predictive Drug Safety: From Structural Alerts to Biological Mechanisms. *Chem Res Toxicol.* **28**, 1875-1887 (2015).
- Yang, C. *et al.* New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. *J Chem Inf Model.* **55**, 510-528 (2015).
- Remez, N. *et al.* The In Vitro Pharmacological Profile of Drugs as a Proxy Indicator of Potential In Vivo Organ Toxicities. *Chem Res Toxicol.* **29**, 637-648 (2016).
- Carrió, P. *et al.* Toward a unifying strategy for the structure-based prediction of toxicological endpoints. *Arch Toxicol.* **90**, 2445-24 (2016).
- Li, T.S. *et al.* A crowdsourcing workflow for extracting chemical-induced disease relations from free text. *Database (Oxford).* pii: baw051 (2016).
- Bravo, À. *et al.* Combining machine learning, crowdsourcing and expert knowledge to detect chemical-induced diseases in text. *Database (Oxford).* pii: baw094 (2016).
- Ravagli, C. *et al.* OntoBrowser: a collaborative tool for curation of ontologies by subject matter experts. *Bioinformatics.* **33**, 148–149 (2017).
- Carbonell, P. *et al.* Hepatotoxicity prediction by systems biology modeling of disturbed metabolic pathways using gene expression data. ALTEX. **34**, 219-234 (2017).

- Cañada, A. *et al*. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. *Nucleic Acids Res*. 2017 May 22. [Epub ahead of print]
- Briggs KA. Is preclinical data sharing the new norm? *Drug Discov Today.* 2016 May 10. [Epub ahead of print]

S2. eTOX-related web sites

SDISC: https://www.cdisc.org/standards

ChEMBL: https://www.ebi.ac.uk/chembl/

Collector: http://phi.upf.edu/collector/

DrugBank: https://www.drugbank.ca/

eTOX: http://www.etoxproject.eu/

eTOXlab: http://phi.imim.es/envoy/

eTOXsys: https://www.lhasalimited.org/products/etoxsys.htm

eTOXsys Dashboard: https://etoxsys.eu/etoxsys.v3-demo-bk/dashboard/

eTOX Hackathon: https://www.youtube.com/watch?v=idvAw8frJqs&feature=youtu.be

Hpath: http://reni.item.fraunhofer.de/hpath/

IMI: https://www.imi.europa.eu/

INHAND: https://www.toxpath.org/inhand.asp

Ontobrowser: https://github.com/Novartis/ontobrowser

Open PHACTS: https://www.openphacts.org/

RepDose: http://fraunhofer-repdose.de/

SEND: https://www.cdisc.org/standards/foundational/send

TransQST: http://transqst.org/