TOXICOLOGICAL SCIENCES, 162(1), 2018, 287–300

doi: 10.1093/toxsci/kfx254
Advance Access Publication Date: November 16, 2017
Research Article

Generating Modeling Data From Repeat-Dose Toxicity Reports

Oriol López-Massaguer,* Kevin Pinto-Gil,* Ferran Sanz,* Alexander Amberg,† Lennart T. Anger,† Manuela Stolte,† Carlo Ravagli,‡ Philippe Marc,‡ and Manuel Pastor*,[1]

*Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Universitat Pompeu Fabra, 08003 Barcelona, Spain; †Sanofi, Preclinical Safety, 65926 Frankfurt am Main, Germany; and ‡Translational Medicine, Novartis Institute for Biomedical Research, CH-4002 Basel, Switzerland

[1]To whom correspondence should be addressed at Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Universitat Pompeu Fabra, C/Dr. Aiguader 88, 08003 Barcelona, Spain. Fax: 00 34 933169559; E-mail: manuel.pastor@upf.edu.

ABSTRACT

Over the past decades, pharmaceutical companies have conducted a large number of high-quality *in vivo* repeat-dose toxicity (RDT) studies for regulatory purposes. As part of the eTOX project, a high number of these studies have been compiled and integrated into a database. This valuable resource can be queried directly, but it can be further exploited to build predictive models. As the studies were originally conducted to investigate the properties of individual compounds, the experimental conditions across the studies are highly heterogeneous. Consequently, the original data required normalization/standardization, filtering, categorization and integration to make possible any data analysis (such as building predictive models). Additionally, the primary objectives of the RDT studies were to identify toxicological findings, most of which do not directly translate to *in vivo* endpoints. This article describes a method to extract datasets containing comparable toxicological properties for a series of compounds amenable for building predictive models. The proposed strategy starts with the normalization of the terms used within the original reports. Then, comparable datasets are extracted from the database by applying filters based on the experimental conditions. Finally, carefully selected profiles of toxicological findings are mapped to endpoints of interest, generating QSAR-like tables. In this work, we describe in detail the strategy and tools used for carrying out these transformations and illustrate its application in a data sample extracted from the eTOX database. The suitability of the resulting tables for developing hazard-predicting models was investigated by building proof-of-concept models for *in vivo* liver endpoints.

Key words: RDT; toxicology databases; *in vivo* data; *in silico* modeling; ontologies.

In vivo repeat dose toxicity (RDT) studies are a compulsory component of the safety assessment of drug candidates carried out for regulatory purposes. The pharmaceutical industry has conducted thousands of RDT studies over the past few decades. Until now, most of the data reported from these RDT studies has not been stored in a structured format or database (even by the companies who conducted the studies). The eTOX project (Sanz *et al.*, 2015), started in 2010, aims to remediate this situation by compiling the reports from RDT studies donated from participating companies, extracting and harmonizing the data therein and storing it in an integrated database. As of May 2016, (version 2016.1), the eTOX consortium has compiled reports

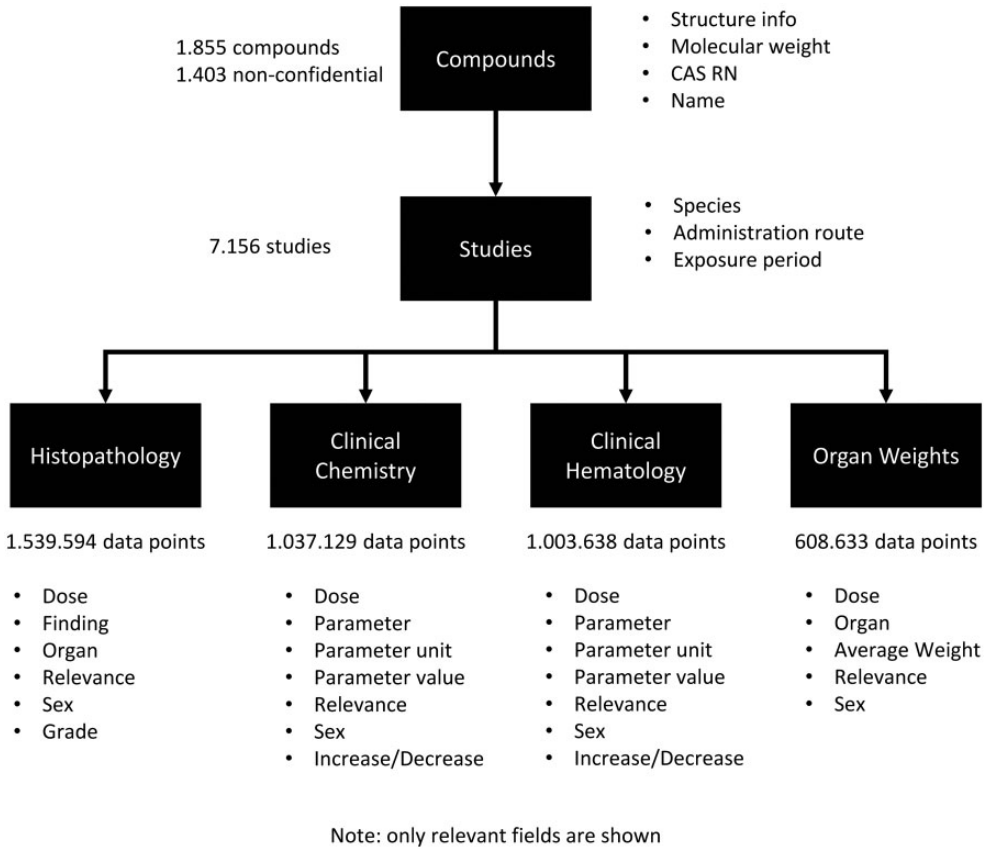

Figure 1. eTOX 2016.1 relational database schema. The schema represents the main tables in the database and their relationships. The main table stores the tested compounds. For every compound we have several studies with diverse designs. For every study the experimental findings are stored in different tables according to its type. Each finding summarizes the experimental observation as well as the dose and study timepoint.

from 7156 studies (1855 compounds) resulting in almost 9 million data points. The eTOX database is a highly valuable resource on its own, which can be queried to extract information about any of the RDT studies. A detailed description of the eTOX database is beyond the scope of this work and can be found here (Cases *et al.*, 2014). However, to summarize the work briefly, study reports were manually curated and entered into a classical normalized relational database scheme (depicted in Figure 1) (Codd, 1970; Date, 1995; Ullman and Widom, 2008). The database contains RDT studies, carried out for a compound with a certain experimental design. The database tables contain observations from diverse experimental domains: clinical chemistry, urine analysis, organ weight, histopathology, etc., which are common to the majority of regulatory oriented RDT studies. Some of these observations (so called "findings") describe an observed effect of a certain compound in a group of animals treated under the same conditions (same dose, administration route and time point). They are labeled as "treatment-related" or "non-treatment-related" based on the original expert toxicological assessment present in the report.

Studies in the eTOX database were carried out by pharmaceutical companies or contract research organizations (CROs) under good laboratory practices conditions. Harmonizing data from many diverse study providers has been a major challenge. Biological results were often captured using inconsistent nomenclatures; hence, the originally reported terms required mapping to target ontologies one of which was developed specifically for the eTOX project (Ravagli *et al.*, 2017).

One of the goals of the eTOX project is to go beyond database Structured Query Language (SQL) queries and explore novel ways of exploiting this valuable resource. In particular, we aim to use the *in vivo* data therein for building computational models able to predict the toxicological properties of drug candidates not present in the database (Sanz *et al.*, 2015). The computational methods used for this purpose work by inferring the properties of new compounds from those observed in the database. The predictive modeling algorithms can be very diverse and the nature of the prediction can be either quantitative or qualitative, but in all the cases the accuracy of the prediction depends on the accuracy of the properties attributed to the drug candidates in the database. This consideration is not trivial, since the diagnostics which we aim to predict (*in vivo* endpoints) do not have a direct correspondence with the descriptive annotations present in the reports and collected in the database (reported findings). For example, if we want to develop a model predicting liver toxicity, this endpoint is not present as such in the database, but as a collection of findings (eg, liver necrosis, increase of transaminases and bilirubin, etc.) which are likely to be observed and noted when the compound produces that type of toxicity. Also, RDT studies are not designed to compare the properties of a series of compounds but to assess the toxicity of a single drug candidate. Indeed, one of the characteristics of RDT studies is that they are designed to observe toxicity and hence the dose range is carefully selected on the basis of preliminary studies to make sure that, at least, the animals exposed to the highest doses exhibit significant toxic effects. For these reasons, not all the compounds were studied at the same

dose or using the same administration route, and there are also differences in species and strains. *In vivo* data are still the "gold standard" in preclinical toxicology, but the aforementioned reasons explain the impossibility of using the reported data in its raw form for generating computational models.

The objective of the work described here, was to develop a methodology addressing the limitations of the raw RDT studies for extracting toxicity scores amenable for the development of predictive models. This methodology was developed within the project eTOX, making use of the project database, and implemented in open source software that can be easily adapted to similar databases. Here we will describe our method, the software and present examples illustrating how our methodology was applied and how the toxicity scores generated from the original studies can be used to build predictive models. This method is applicable to any other database collecting RDT reports, such as ToxRefDB (Martin *et al.*, 2009), or in-house pharma companies databases.

MATERIALS AND METHODS

Storage of eTOX Database

To facilitate data processing, the eTOX database (2016.1 version) was downloaded and stored in a local PostgreSQL database (PostgreSQL, 2016; Stonebraker *et al.*, 1986) containing only tables relevant for our purposes. The schema of this database is very close to the simplified view of the eTOX 2016.1 database shown in Figure 1. The data import was carried out using a set of extraction, transformation and loading scripts (Kimball *et al.*, 2008).

Ontologies Storage

The ontologies used to annotate histopathology findings (ie, the anatomy ontology and histopathology ontology) were stored centrally within the eTOX project using a tool called OntoBrowser (Ravagli *et al.*, 2017). The anatomy ontology is an extended version of the Adult Mouse Anatomy (Hayamizu *et al.*, 2005; Mouse Anatomy, 2017) while the histopathology ontology was built by the eTOX Consortium and aligned to INHAND (Keenan *et al.*, 2015). A massive curation effort took place to map all verbatim terms used in reports to preferred terms from the ontologies. These mapping were generated and stored into the OntoBrowser. Ontologies were extracted from the eTOX central repository (Ravagli *et al.*, 2017) and stored in parent-child relational form in the PostgreSQL database. The type of relationship between concepts (eg, is_a, part_of, etc.) and all the synonyms of the preferred terms within the ontologies were also captured within the same database.

Data Extraction

All data extracts from the eTOX database were carried out using an in-house developed tool written in Scala (Odersky *et al.*, 2004) and Slick (Typesafe Slick, functional relational mapping for Scala). Slick is a domain-specific language (Mernik *et al.*, 2005) that provides a mechanism directly within Scala to execute queries against relational databases following the ideas of Kleisli (Wong, 2000). This enabled us to develop a very flexible query building tool incorporating the complex and diverse filtering required for extracting data from the database. For example, we were able to define filtering criteria for studies (eg, based on sex, exposure period, and/or administration routes) in addition to the observation properties (eg, type of observation, relevance, organ, etc.).

This tool implements a novel query reformulation method (Necib and Freytag, 2004) to extract information from the relational database. Essentially, the query reformulation rewrites a conventional SQL query based on a single value condition (eg, organ = "liver") to a query that uses all the synonym terms as well as child terms in the ontology tree. This reformulation method is based on ontology reasoning but instead of implementing it using standard Web Ontology language (OWL) reasoners (Baader *et al.*, 2010) we used PostgreSQL recursive queries (Garcia-Molina *et al.*, 2008). We chose this approach because it allows to integrate the inference and the query processes. Technically, we computed the transitive closure of "part_of" relationships. For example, by recursively traversing the "part_of" relation starting from "liver" we obtain all the anatomic components that are considered parts and subparts of the liver in our anatomy ontology. With respect to classical OWL reasoners, this method has the advantage that it integrates reasoning and querying in a single step and a single tool. The application of this query strategy enriches the raw information stored in the database by augmenting it through logical reasoning.

Model Building

The quantitative and qualitative scorings for diverse *in vivo* endpoints (described in the "Results" section) have been used to build predictive models which represent their association with different descriptors of the compound structure. For the qualitative scorings, the models aim to predict if a new compound is positive or negative. In the case of the quantitative scorings, the model predicts the lowest dose at which an endpoint would be observed.

The model building process started by selecting a series of small molecules annotated with both qualitative and quantitative scorings for the 3 following endpoints: (see "Case-study application: liver toxicity" section), "degenerative lesions" (DEG), "inflammatory liver changes" (INF), and "nonneoplasic proliferative lesions" (PRO). These endpoints were observed in 332, 258, and 246 compounds respectively. Two-dimensional structures were extracted from the eTOX database in SMILES format and they were then normalized using standardizer (Atkinson, 2014). Then, reasonable 3D structures were obtained using CORINA (Sadowski *et al.*, 1994). The ionization state of every compound was adjusted to pH 7.4 using MoKa (Milletti *et al.*, 2007; Molecular Discovery, 2017a). The series of compounds was then divided randomly into a training series (for building the models), which contained 70% of the compounds and a test series (for external validation) containing the remaining 30%. The structures in the training series were used to compute Adriana (Molecular Networks, 2017) and GRIND2 (Durán *et al.*, 2009; Molecular Discovery, 2017b; Pastor *et al.*, 2000) molecular descriptors using respectively AdrianaCode (Molecular Networks, 2017) and Pentacle (Molecular Discovery, 2017b) software. This procedure produced a matrix of descriptors (**X**) which was submitted to a panel of machine learning methods described below (partial least square-regression [PLS-R], PLS-discriminant analysis (PLS-DA), and random forest [RF]) to obtain mathematical functions describing the relationship between them and the toxicity scorings, suitable for predicting the properties of new compounds.

Partial least squares. We used an in-house implementation of PLS-R (Wold *et al.*, 2001) and PLS-DA (Wold *et al.*, 2001) using Nonlinear Iterative Partial Least Squares (NIPALS) algorithm. For the PLS-DA, the cutoff value used to discriminate positive from negative objects was adjusted automatically during the building

step to obtain an optimal balance between the model sensitivity and specificity (characterized by the minimum difference between both parameters), in the training series. We applied the NIPALS algorithm to extract a maximum of 7 latent variables, retaining in the final model the dimensionality producing best predictive quality according to leave one out (LOO) cross-validation. The models were further refined by using Fractional Factorial Design variables selection (Baroni *et al.*, 1993). ADAN (Carrió *et al.*, 2014) methodology was applied for assessing the applicability domain of predicted compounds and to provide pseudo 95% CIs, as described in Carrió *et al.* (2014).

Random forest. We used the RF-regressor and RF-classifier (RF-C) available in the scikit-learn library (Pedregosa *et al.*, 2011). The number of trees (n_estimators), and max features (features) were adjusted using a grid search algorithm to obtain optimum values according to out of bag criterion. Additionally, to obtain a comparable assessment of the models' predictive quality, we carried out LOO cross-validation.

The performance of the qualitative models has been assessed computing the sensitivity (sen), specificity (spe) and Matthews correlation coefficient (MCC), as described in the equations below:

$$\text{sen} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

$$\text{spe} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{2}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \tag{3}$$

The goodness-of-fit of the quantitative models was assessed using the determination coefficient (r^2). Their predictive performance was assessed using LOO cross-validation and quantified using cross-validated determination coefficient (q^2) and the Standard Deviation Error of the Prediction (SDEP).

Finally, all models have been further validated by predicting the test series and computing the external sensitivity, specificity and MCC parameters for the qualitative models and q^2 and SDEP for the quantitative ones.

RESULTS

Overview

The methodology described here generates data suitable to model hazard for *in vivo* endpoints from a collection of *in vivo* RDT studies. The original data is processed to produce hazard scores assigning a single value to each compound that characterizes its properties with respect to a certain *in vivo* endpoint. Ideally, such scores must be quantitative and the scale must represent the compound hazard in "point of departure" (PoD) terms, the dose from which the toxic effect starts to appear. The values must be comparable between compounds allowing to rank them from lower to higher hazard.

This work does not aim to produce risk scores, because the data available in the eTOX database is too limited for modeling the substance exposure without introducing toxicokinetic (TK) information from external sources. Also, for similar reasons, the scores will express hazard in terms of substance dose and not absolute concentration in the target organ.

The proposed approach involves 3 steps: (1) study filtering, (2) finding aggregation, and (3) score generation, which will be described in detail in the following sections. The quality of the scores obtained in a case-study application will be analyzed in terms of their suitability for modeling (eg, value distribution, number of positives, etc.) but also by analyzing the predictive quality of example models developed from them.

All these steps were implemented in a set of scripts written in Scala (Odersky *et al.*, 2004) which generate the complex SQL queries required to extract the data. Some of these scripts include a web-based graphical user interface (GUI) that facilitates their use. All this software and the associated documentation is distributed as open source under license GNU v3.0 (Free Software Foundation, 2007) and can be downloaded from the URLs included in the following references (Lopez-Massaguer, 2017a,b). The commands used to generate the examples in this article were included in section source code 1 (Ontology based data extraction) and source code 2 (Scoring computation scripts) of the Supplementary Material.

Study Filtering

The eTOX database integrates studies conducted on different species/strains with various compound administration routes and diverse dosing durations. Figure 2 illustrates the diversity of data collected and the number of compounds and data points present in the database.

The generation of comparable scorings requires starting from a set of comparable studies with a reasonable degree of similarity with respect to the aforementioned design characteristics. However, any study consistency criteria applied here will impose a limit to the number of substances and studies used for generating the scores, and therefore a suitable compromise is required. For example, using all the studies and substances available in the eTOX database (7156 studies and 1855 compounds for the 2016.1 version) will be unacceptable, since this will merge studies from rat, mouse, dog, monkey, etc. (see Figure 2) as well as studies of different duration and route of administration. Conversely, if we restrict our analysis to studies carried out only with Wistar rats, 20-day treatment duration, oral gavage, only 6 out of the 7156 studies would be selected (4 out of 1855 compounds).

Our approach uses an iterative process, relaxing slightly the selection criteria and checking the effect on the number of studies/compounds until obtaining a reasonable compromise between the consistency and the number of the studies selected. In the previous example, the use of a relaxed criteria permitting a range of related species (rodents: rats and mice of any strain), routes (oral: feeding and gavage) and duration (approximately 4 weeks, from 21 to 32 days) allowed selecting 883 out of the 7156 studies (597 out of 1855 compounds). The effect of the 2 filtering strategies is illustrated in the Tables 1–3.

The use of such iterative approach was made possible by an interactive data extraction tool developed in house. Conceptually, the tool is based on the On-Line Analytical Processing (OLAP) multidimensional model (Agrawal *et al.*, 1997; Gray *et al.*, 1997); the original relational database has been transformed in a multidimensional database (Figure 3) where the "facts" are the findings and the "dimensions" are the characteristics of the study design and the finding, as listed in Table 4. The transformation was done to adapt the original database model to a data extraction oriented model, based in the OLAP principles.

Based on this conceptual model, we can filter the studies by defining 1 or several acceptable values for fields containing qualitative variables (eg, species) and/or a range of acceptable values for fields containing quantitative data (eg, exposure

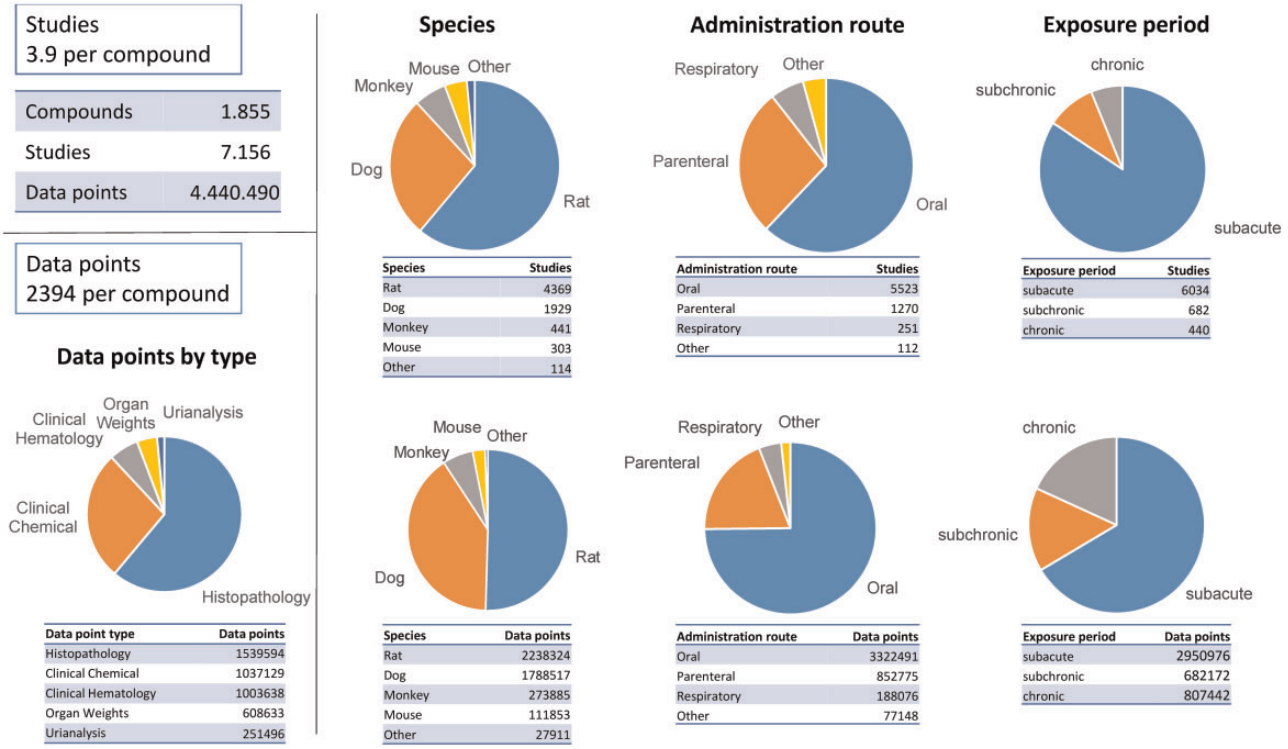

Figure 2. Database statistics. This figure depicts some general database statistics: number of compound, studies and findings. It also shows the distribution of studies and findings by species, administration route and exposure period as well as the distribution of findings by finding type, and the ratio of studies per compound and findings per compound.

Table 1. Filtering Conditions Used

	Narrow	Broad
Species	Rat	Rat, mouse
Administration route	Oral	Oral, oral gavage
Exposure period	20 days	20–32 days

Table 2. Summary of Studies/Compounds Selected

	Narrow	Broad
Studies	6	883
Compounds	4	596

Table 3. Summary of Observations Selected

Finding Type	Narrow	Broad
Clinical chemistry (general)	595	119 860
Clinical chemistry (haematology)	668	111 920
Histopathology	1576	244 721
Organ weights	8	126 717
Clinical chemistry (urine)	0	36337

period). Table 5 describes the characteristics which can be used and shows some examples.

In terms of OLAP model, this data operation (filtering the database by choosing several values for 1 dimension) is equivalent to "dice" the multidimensional data cube (Figure 4).

In our tool, the values of the filtering parameters can be selected using a web-based GUI or a command line interface. In either case, the query parameters are translated internally to SQL queries that apply the filtering.

Most of the studies and finding characteristics listed in Tables 4 and 5 are selfexplicative (eg, species, route, and organ) but the "relevance" field deserves a separate explanation. Only findings which were declared by the responsible toxicologist to be caused by the drug candidate were labeled as "treatment-related". Therefore, this field introduces expert judgment adding valuable insight in the database, even if it is always a subjective appreciation and cannot be considered infallible.

One of the sources of complexity in this filtering is the need to consider the diverse terms which can be used to describe anatomic structures or histopathology findings. For example, in the case of selecting findings by the affected organ (eg, liver), the end-user typically wants to obtain findings affecting any part of the liver (eg, liver lobule, liver acinus, portal triad, etc.). Incorporating this anatomical understanding into the queries required the use of an anatomic ontology capturing the many different terms used in the report to describe the anatomic structures as well as their relationships (eg, "is part of"). This knowledge is used internally by our tool to expand the filtering. For example, the term "Liver" is searched in the ontology and all concepts connected by "is part of" are added to the collection of queried terms recursively (see an example in Figure 5A). The complete list of "Liver" connected terms can be seen in the Supplementary Table 1.

A similar, but more complex situation is the filtering applied to histopathology findings (eg, necrosis). As in the previous case, the query is expanded by recursively exploring a suitable ontology term; in this case using "subclass of" relationships (see an example in Figure 5B). The final result is a collection of terms describing the queried finding with different levels of detail

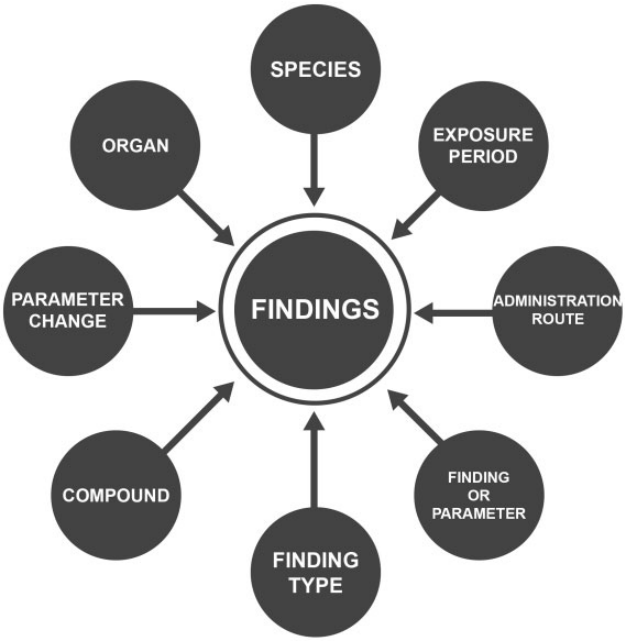

Figure 3. OLAP schema. Schema of the multidimensional data model used to build the data extraction mechanism. Findings are modeled as facts in a multi-dimensional space, located at the center and connected to the dimensions around, describing 8 different properties: species, organ, parameter change, compound, finding type, finding/parameter, administration route and exposure period.

Table 4. OLAP model dimensions.

Dimension	Example
Species	Rat, mouse
Administration route	Gavage, intra venous
Exposure period	(number of days)
Finding class	Clinical chemistry, histopathology, organ weights
Finding type	In case of qualitative observations like histopathology findings: necrosis, inflammation
Parameter	In case of quantitative parameters like clinical chemistry findings: ALT, AST, Bilirubin
Organ	Liver, spleen
Treatment relevance	Treatment-related/non treatment-related
Change	Change in quantitative values: Increased/decreased

(eg, necrosis zonal, necrosis zonal centrilobular, necrosis zonal periportal), all of which are used to query the database. The complete list can be seen in the Supplementary Table 2.

Finding Aggregation

The result of the filtering step is a collection of several findings observed in comparable studies, for diverse compounds at diverse doses. This collection is represented in the table at the left-hand side of Figure 6 and often contains several rows per compound and finding, indicating that for a certain compound the same finding has been observed at different doses, in some cases in different studies.

The findings extracted from the database often contain a nonhomogeneous level of detail, for this reason, the raw collection of findings is preprocessed using the same ontology described above to obtain parent findings, thus facilitating posterior aggregation. For example, if the extraction obtains "perilobular necrosis", the parent finding "necrosis" is also assigned to the same compound.

The next step of the process is to collapse all the aforementioned rows into a single line representing a single compound-finding combination. This can be done in 2 different ways: quantitative, by recording the lowest dose at which the finding was observed, or qualitative, by simply recording that the finding was observed in any of the studies. The aspect of the tables obtained is shown in the middle of Figure 6. In the case of the dose, it can be interpreted as a LOEL (lowest observed effect level), even if it is not strictly correct, since this dose can be obtained from a collection of studies.

In any case, the resulting table still contains more than 1 row per compound, and to obtain a QSAR-like table, we need to apply a pivoting transform, assigning a column to each finding and include in the compound-finding intersection either the LOEL value or a binary indicator (see the rightmost column in Figure 6). At this point it must be stressed the importance of the previously described ontology preprocessing (rollup) of the extracted findings. This operation effectively maps detailed endpoints to higher-level parents.

Scoring Generation

The tables produced by the aggregation described earlier contain a single row per compound but many columns describing individual findings. Each of these findings represents an "observable manifestation" of an effect (endpoint) produced by the compound, which is the real object of the study, and the biological property that we aim to predict using our computational models.

Diagnostics or endpoints like "liver inflammation" or "cholestasis" are not described in the report as such. For this reason, our method proposes to infer these endpoints from a profile of observed findings. All the findings considered to be associated to the endpoints are combined using a logic OR

Table 5. Filtering capabilities, including parameters and examples.

Filter Level	Filter	Filter Parameter	Example	Only for Table[a]
Study	Species	One or more species of the study	Rat or mouse	
Study	Administration route	One or more administration routes of the study	Oral or oral gavage	
Study	Exposure periods	A range of days for filtering the exposure period of the study	From 20 to 32 days	
Findings	Relevance of findings	Include or not only treatment related findings	Only treatment-related findings	
Findings	Organ	One or more organs of interest	Liver	Histopathology
Findings	Finding	One or more type of findings	Necrosis	Histopathology
Findings	Clinical chemistry parameter	One or more Parameter	ALT or AST	Clinical chemistry
Findings	Change in clinical chemistry parameter	Increase or Decrease	Increase	Clinical chemistry

[a]These characteristics of the findings are exclusive of the tables listed here.

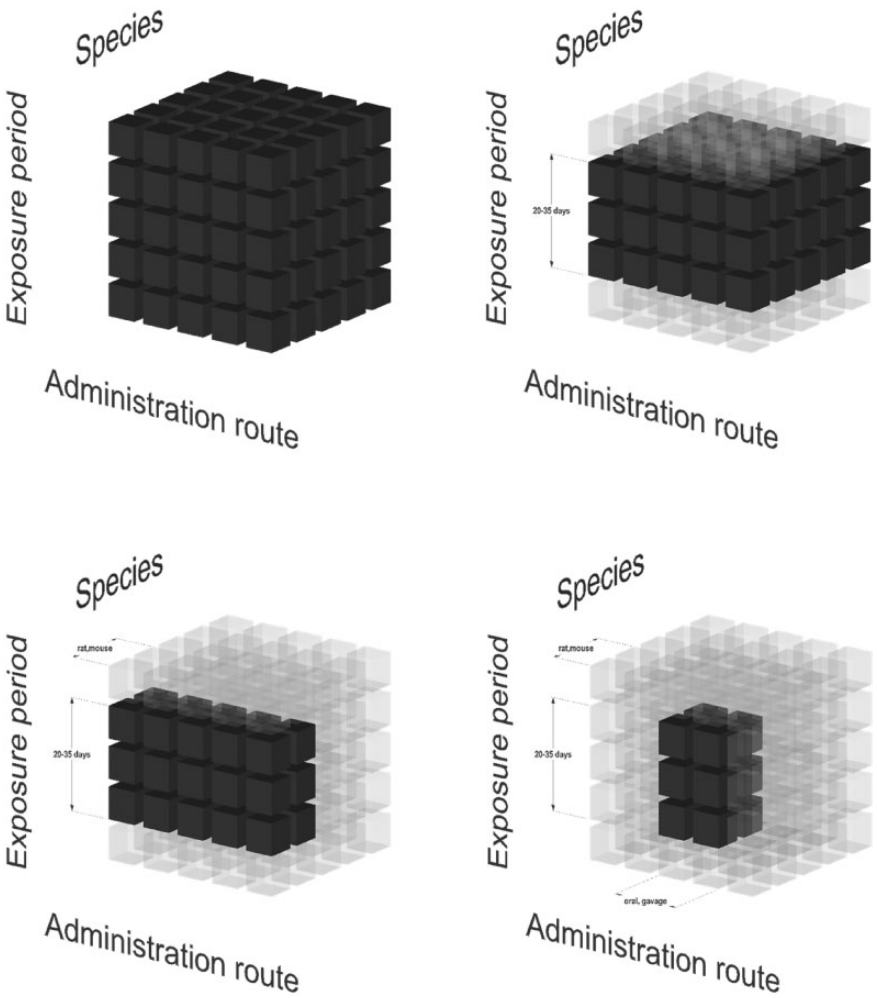

Figure 4. Slicing OLAP cube. Schematic representation of the database filtering as the slicing of a data cube, where the findings are located in a 3D space, with axis representing exposure, administration route and species. The filtering process can be graphically represented as a progressive slicing of this multidimensional data cube, selecting only the findings enclosed within the ranges of values defined for every axis. This representation is a simplification, since the actual data cube is 8-dimensional.

operation, which means that compounds for which any of the findings in the profile is observed will be assumed to produce the endpoint. For example, if we consider that "degeneration" is strongly associated with findings like "amyloidosis" and "mineralization", the observation of any of these findings in a compound will be used to consider it positive for the endpoint. To avoid false positives, it is advisable to consider only "treatment related" findings, as explained in the Filtering section. It is important to stress that the definition of the finding profiles associated to the toxicity endpoints considered in this

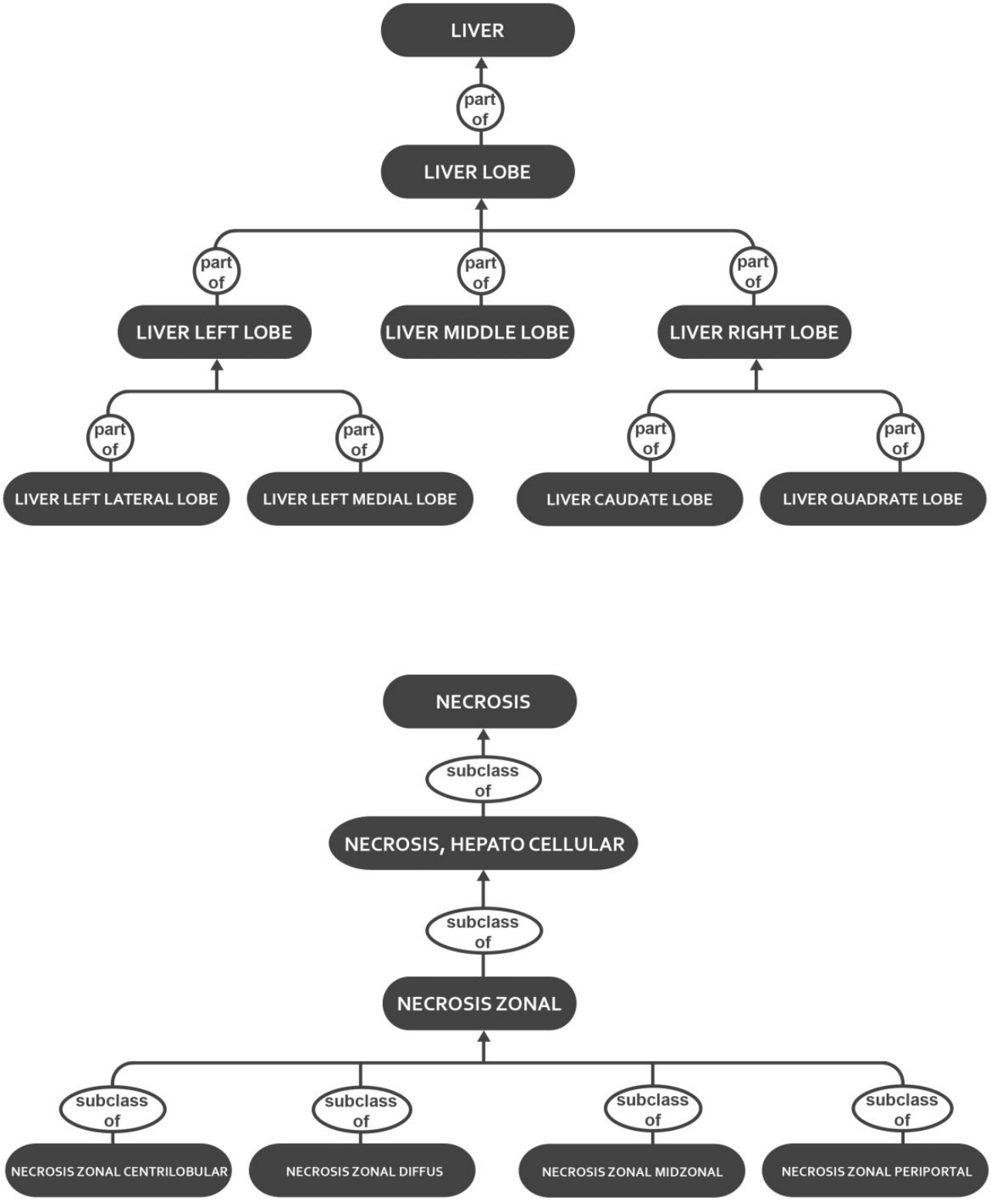

Figure 5. Ontologies. A, Fragment of the Organ ontology showing a subset of concepts related to the "Liver" concept by the predicate "part_of." B, Fragment of the Histopathology ontology showing a subset of the concepts of the ontology related to "Necrosis" concept by the predicate "subclass_of."

study was carried out by expert toxicologists, without the use of any statistical method. In this study, in line with a previous work (Mulliner *et al.*, 2016), the association between findings and endpoints was described using a hierarchical structure, defining a first level of finding clusters, associated with specific endpoints (eg, necrosis) and a second level association of such endpoints to describe higher level *in vivo* endpoints (eg, degenerative lesions). A complete list of the findings and their first

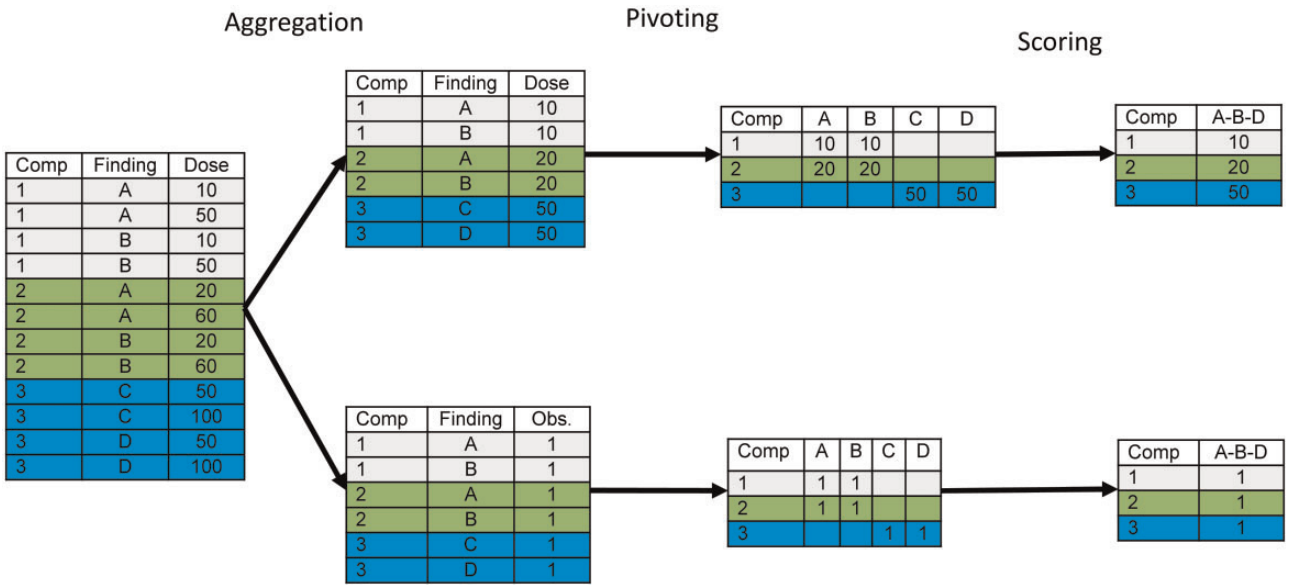

Figure 6. Aggregation. Schema of the method proposed for calculating quantitative (upper path) and qualitative (lower path) hazard scores. Starting from a collection of findings, these are aggregated by compound producing only 1 line per compound and finding, containing the lowest dose in the case of quantitative scores and a binary variable for the qualitative ones. This table is pivoted, producing a single line per compound and separating the findings in columns. Finally, we use a profile of findings known to be associated to the endpoint (A-B-D in this case) to collapse these values into a single compound score.

and second level association profiles was included in Supplementary Tables 3 and 4.

In this study, we applied "finding profiles" to generate qualitative and quantitative scorings. The qualitative scoring only assigns to the compound a positive value if any of the findings of the profile has been observed. The quantitative scoring assigns to the compound the lowest LOEL of these findings. The LOEL values, originally in mg/kg/day, were transformed to molar units by dividing them by the molecular weight. This value is scaled using a minus logarithmic transform to obtain values following a growing scale: the more positive the scoring is the lower the dose is at which the first sign of the toxicity was observed. The logarithmic transform make these scorings also amenable for QSAR according to the extrathermodynamic principle (Leffler and Grunwald, 1963).

The result of this process is a list that only contains "positive compounds", annotated to at least one of the findings in the profile. Model building methods also require the inclusion of a suitable collection of "negative compounds". This requires the identification of compounds tested under similar conditions to the positive but for which no relevant finding was observed. We must stress that selecting compounds for which nothing was noted is not enough, since this will merge untested and negative compounds. For this reason, the selection must require a minimum dose tested, and the definition of any other characteristic that makes a compound non-eligible as "negative", typically if the findings are treatment-related or not, and the tables/organs affected. Quantitative models require assigning a toxicity score also to the negative compounds. Any low value would represent their distinction from the positives. However, to incorporate the effect of the tested doses, we choose to assign them a score representing a dose 10 times higher than the maximum value tested. This means that, for modeling purposes, we are arbitrarily assuming that all compounds are toxic, but in the case of the negatives, at a dose higher than the highest dose tested. It must be stressed that this approach requires the application of very strict selection criteria for the negative compounds, like those described earlier. Otherwise we risk assigning similar scores to negative compounds than to toxic, but poorly bioavailable ones.

Case-Study Application: Liver Toxicity

As an example of the whole procedure, we present here a case-study application of the method for obtaining qualitative and quantitative scores for 3 liver toxicity endpoints: "nonneoplasic proliferative lesions" (PRO), "inflammatory liver changes" (INF), and "degenerative lesions" (DEG). For this, first all available primary treatment-related and standardized microscopic liver findings from the studies were clustered into groups of similar terms from a pathological point of view (first level cluster). For example, for the first level cluster term "necrosis" the following primary terms were aggregated: "necrosis"; "necrosis, fibrinoid"; "necrosis, focal/multifocal"; "necrosis, hepatocellular"; "necrosis, zonal"; "necrosis, zonal, centrilobular"; "necrosis, zonal, midzonal"; "necrosis, zonal, periportal"; "single cell necrosis"; "single cell necrosis, epithelial". Then, different first level clusters were aggregated into groups of similar second level cluster, also from a pathological point of view. For example, for the second level cluster "degenerative lesions" the following first level cluster were aggregated: "necrosis"; "apoptosis"; "vacuolation"; "fatty changes"; etc.

The first step was to apply sensible filters to extract a collection of comparable studies. Table 6 lists the criteria applied, aimed to extract rat studies of any duration and oral administration (excluding dietary). We obtained 2226 studies of these characteristics. From the 1855 compounds present in the database, only 934 appear in any of these studies and incorporate finding annotations. This list of compounds constitutes the base chemical universe (BCU) of our study, and any further filtering is restricted to only these compounds.

For the compounds of the BCU we extracted from the histopathology table treatment-related findings in which liver is the organ affected. At this point it is worth mentioning the effect of the query expansion using the organ ontology (to add findings

assigned to parts of the liver) and the histopathology ontology (to add parent findings): the raw number of findings was 6207, but the use of the anatomy ontology increased this number to 6245 (38 more, 0.6% increase) and the histopathology ontology to 7292 (1085 more, 17% increase).

The second step was to aggregate all these findings per compound. As illustrated in Figure 6 the compounds were associated with the corresponding findings either using a binary indicator or the lowest dose at which this finding was observed in any study, thus producing a qualitative and a quantitative table with a single row per compound and columns representing every finding observed.

The final step was the generation of the scorings for the 3 considered endpoints, using the finding profiles shown in Supplementary Tables 3 and 4 generated by expert toxicologists. We computed 2 kinds of scorings (quantitative and qualitative) as described in the Materials and Methods section: qualitative (logical OR of the findings observed) and qualitative (minimal dose of the observed findings transformed into logarithmic scale). The positive compounds obtained are listed in Table 7. For this particular case-study negative compounds were defined as (1) substances included in the BCU, (2) tested at least at a dose higher or equal to 1000 mg/kg (252 compounds), and (3) with no observed treatment-related, liver-related histopathology findings. These criteria guarantee that the negative compounds have been tested (1) in conditions comparable with the positives one, (2) at doses high enough to avoid selection of untested compounds, and (3) excluding compounds with any liver-related finding, even unrelated to those in the scoring profile. These criteria can be seen as restrictive, but even so the number of negative compounds obtained is already larger than the positive ones (see Table 7), and there is no benefit in obtaining a larger series.

A closer inspection to the scores shows that, for most positive compounds more than one of the findings of the profile has been observed. The median number of findings is 3 for DEG and INF and 4 for PRO. Bar charts describing the full distribution of values have been included as supplementary information (Supplementary Figure 3).

Table 6. Hepatotoxicity filtering criteria used for the case-study application.

Filter Level	Filter	Value
Study	Species	Rat
Study	Administration route	Intraesophageal Intragastric Nasogastric Oral Oral gavage Oropharyngeal
Study	Exposure periods	Any
Findings	Relevance of findings	Treatment related
Findings	Organ	Liver
Findings	Finding	Any

Regarding the quantitative scorings, their means and standard deviations have been included in Table 7 and histograms with the distribution of the values were also included in Supplementary Figure 3. The distribution is multimodal, probably due to the use of discrete doses. As expected, the values cover a relatively narrow range of 3–4 log units, with SDs under 1 log unit. The values represent pseudo-LOEL doses in mg/kg/day, and not potency expressed as molar concentration, as it is usual in QSAR modeling. All in all, these circumstances (narrow range, no incorporation of pharmacokinetics) can be expected to limit the quality of quantitative models, but in the next section we present some preliminary models that show how, in spite of the limitations, the obtained scores represent a reasonable starting point to describe the toxicity of a collection of compounds in a comparable way.

Finally, we have run a comparative analysis of the quantitative LOEL with other PoD reported in the eTOX database. Unfortunately, these data are only present for a few number of the drug candidates in the BCU (572 NOAEL, 301 NOEL, and 14 LOAEL) and the database does not specify the responsible endpoint, but it still could be expected a rough correlation between these values and our pseudo-LOEL. The results are depicted in Figure 7 where we plot the pseudo-LOEL for all LIVER endpoints compared with the NOAEL (maximum NOAEL), expressed as mol/kg/day.

Remarkably, the pseudo-LOEL values obtained from liver-related findings exhibit the expected correlation. Some of the reported LOAEL and NOAEL are lower than our quantitative LOEL, probably representing endpoints not considered in our analysis. Only for a few compounds the reported PoDs are slightly higher and only in 1 case the difference is larger than 1 logarithmic unit. These deviations can be explained by the fact that not all effects in our LOEL represent adversity. In any case, the scatterplot clearly shows the existence of a relationship, further justifying the relevance and usefulness of the proposed quantitative scoring.

Model Building and Results

The scorings obtained were used to build models for the 3 selected endpoints as described in the Materials and Methods section and their quality was evaluated using both LOO cross-validation and external validation. The results for the qualitative and quantitative scorings were summarized in Supplementary Tables 5–11.

A detailed analysis of the effect of the different molecular descriptors and machine-learning methods on the model quality is out of the scope of this work. The models were shown here only to show that the hazard scores generated by our method can be used for modeling, and therefore we will focus on the description of the best qualitative and quantitative models (highlighted in green in Supplementary Tables 5–11).

Qualitative Scores

The choice of the best model was based on the values of sensitivity and specificity obtained for the external prediction,

Table 7. Distribution of values for qualitative and quantitative hepatotoxicity scorings.

Endpoint	Number of Negatives	Number of Positives	Median Number of Findings	Mean Quant. Scorings	SD Quant. Scorings
DEG	168	164	3	3.76	0.84
INF	164	94	3	3.88	0.92
PRO	164	82	4	3.84	0.95

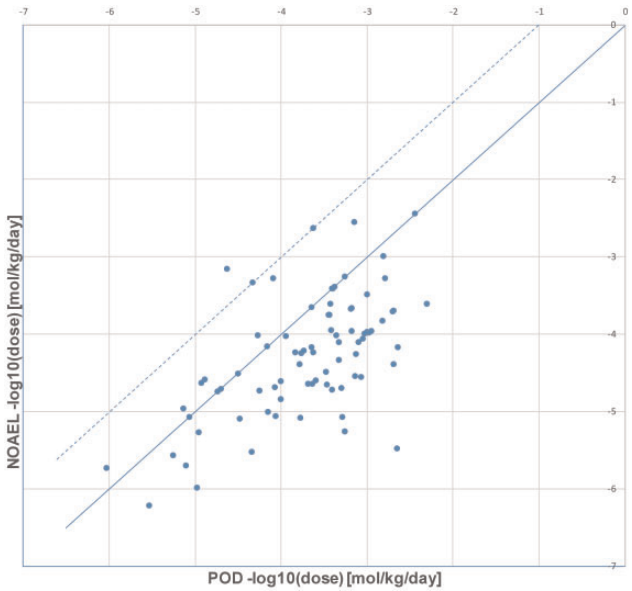

Figure 7. NOAEL vs pseudo-LOEL. Scatter plot representing the NOAEL reported in the studies with our computed pseudo-LOEL values in logarithmic scale as mol/kg/day. If the proposed simplifications are reasonable, most values must be under the diagonal line. This happens for the vast majority of cases, with a few exceptions showing deviations no larger than a logarithmic unit and a single compound with slightly larger deviations.

represented in Figure 8A. The best models for the three endpoints were obtained with Adriana descriptors. Regarding the modeling technique, RF-C produced better results for DEG and INF, while PLS-DA produced slightly better results for PRO.

For the DEG endpoint, the cross-validation results were acceptable (spe = 0.55, sen = 0.68) and similar to the results obtained using the external series (spe = 0.59, sen = 0.67). For the INF endpoint we observed a similar situation, even if the results of the predictive quality assessment were slightly better, both in cross-validation (spe = 0.84, sen = 0.44) and for the external series (spe = 0.76, sen = 0.54). In these models, the use of RF-C produced a perfect fitting of the training series (spe = 1.0, sen = 1.0) which has no value for evaluating the model quality. Conversely, for the PRO endpoint, built using PLS-DA, the fitting obtained for the training series was not so good (spe = 0.75, sen = 0.75) and these quality parameters are closer to those obtained in cross-validation (spe = 0.69, sen = 0.70) and for the external series (spe = 0.62, sen = 0.50).

Assuming that the external prediction produces the strictest evaluation of the predictive performance, we can see that we obtained values of spe and sen over 0.5 in all cases, with a range between 0.5 (PRO sen) and 0.76 (INF spe). It is noteworthy that the unbalanced datasets (Table 7) present in INF and PRO endpoints produced less sensitive models (Figure 8A) while the more balanced dataset (DEG) (Figure 8A) yields a better spe-sen compromise and higher sensitivity. In general, the quality of the models is not excellent, but is surprisingly good if we consider the complexity of the *in vivo* endpoints based on many different mechanism of liver toxicity that are being described and the simplicity of the model approach described here.

Quantitative Scores

We selected as the best model the one with the lowest SDEP for the external prediction. PLS-R produced the best results in all instances, but for DEG and INF the best molecular descriptors were the Adriana ones, while for PRO Pentacle worked better.

In general, the quantitative models were poor, with goodness-of-fit statistics ranging between 0.26 and 0.58. We only obtained a positive value for the external q^2 (0.07) for the PRO endpoint, the only represented in Figure 8B. However, the scatterplots suggest the presence of a linear trend, particularly for the training series and confirm the feasibility of using the proposed scores for quantitative modeling of *in vivo* endpoints.

DISCUSSION

The above-mentioned example illustrates how the proposed hazard scores can be used for modeling, even if they are imperfect and suffer from different limitations that need to be discussed in detail to understand their potential applications.

The first step is to select a subset of studies carried out in comparable conditions (with respect to species, treatment duration, administration route, finding type, etc.). This selection is always a compromise between selecting a very homogeneous set of studies (but with a limited number of compounds fulfilling the selection criteria) and relaxing the selection criteria for obtaining larger series (but including studies carried out in very different conditions). Our method does not impose any constraint and the end user is offered a flexible filtering tool which can be applied interactively, exploring the effect of different selection setting on the number of compounds selected. This leaves to the end-user the responsibility of calibrating the effect that merging studies obtained in different species, administration routes or duration might have on the quality of the scores.

RDT studies are designed to observe toxicity and, therefore, at the highest tested dose, most studies report some toxic effect. But some of these might be not representative of the drug candidate chronic toxicity, since at very high dose the animals often develop nonspecific toxicity that produces alterations at multiple levels. However, in our study, this contamination is partially mitigated by the use of the "treatment related" flag, based on expert toxicological judgment. Even if it is always a subjective appreciation and cannot be considered infallible, in many cases such nonspecific or secondary effects will not be labeled as related with the drug candidate by the expert signing the report. Moreover, in the proposed quantitative scores, findings observed at a high dose will correspond to the lowest scoring values, very close to the values assigned to the compound for which the finding was not observed.

Another important consideration, briefly discussed in the Results section, is the selection of negative compounds. Choosing substances for which the relevant findings have not been observed is not sufficient, because in some cases the doses tested could have been too low to cause toxic effects. For this reason, it is important to select compounds tested at doses high enough to guarantee testing conditions comparable to the compounds in the positive series. It is also important to use ontologies to normalize findings. Otherwise, one risk is to classify a compound as negative for certain finding (eg, necrosis) for which an identical finding was observed but described in more detail (eg, necrosis zonal centrilobular).

Regarding the quantitative scorings, it must be noted that these are expressed as dose and not as the organ specific *in vivo* drug exposure. This simplification neglects TK effects like the compound absorption, bioavailability, renal and hepatic clearance as well as protein binding, to name only a few. The incorporation of these effects would be possible with diverse degrees of accuracy, depending on the availability of TK parameters.

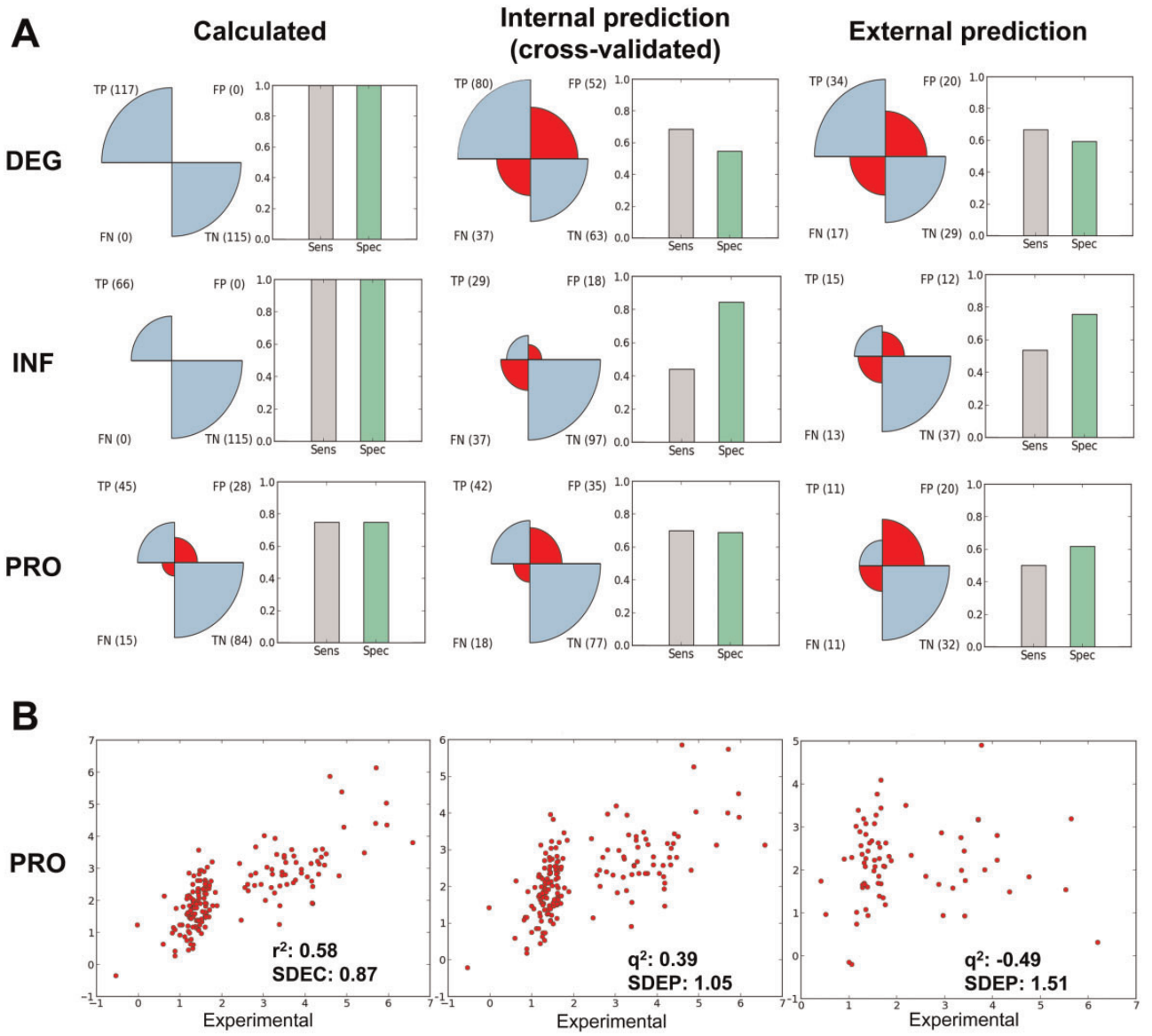

Figure 8. Best qualitative and quantitative models. Calculated values correspond to the values assigned by the model to the compounds in the training series. Internal predictions were obtained using LOO cross validation. External predictions represent real predictions of 30% of the original series, not used for model building. A, Representation of the confusion matrix, sensitivity and specificity for the best qualitative models obtained for the different endpoints. B, Scatterplot showing the best quantitative model, according to the lowest SDEP for external prediction, for the PRO endpoint.

In our case, limiting the analysis to compounds for which TK parameters were available (like maximum plasma concentration, area under the curve, etc.) would have reduced the size of the series to unacceptable levels.

The models presented here illustrate how the proposed scores are amenable for modeling, even if the predictive quality is limited and it is doubtful that, in its present form, they can be used for predictive purposes. In order to obtain better models diverse paths can be explored. First, we can try to overcome the score limitations listed above, refining the way the data is extracted and the criteria used to associate a compound with a given endpoint. The criteria used to assign negative compounds are particularly important, as well as the incorporation of TK parameters for the quantitative scores. Another strategy is using these scores in more sophisticated modeling approaches, representing better the complexity of the endpoints, most representing a mixture of diverse toxicity mechanisms. An example of this strategy was also published by our group (Carbonell *et al.*, 2017) where we applied a systems biology approach to model *in vivo* liver toxicity using an earlier version of the scoring method described here. Finally, recent advances in machine learning methods (Angermueller *et al.*, 2016) and the use of deep learning techniques (Mayr *et al.*, 2016) suggest that we can still improve the predictive quality of models obtained using our proposed hazard scoring.

CONCLUSIONS

We have described a practical method that can be used to transform the information present in databases from RDT reports into hazard scorings for *in vivo* endpoints. The characteristics of these scorings in terms of positive-negative balance and value distributions make them challenging but suitable for building predictive models, as was illustrated in the results section.

The proposed procedure involves a mixture of standard database handling techniques (filtering, aggregation, pivoting) with other, more sophisticated, involving the use of ontologies plus reasoning. In this study, all these operations were carried out very simply, with the use of ad hoc developed software tools, which are accessible as open source under GPL license. They were specifically developed to be used with the eTOX database but, given that we release the source code, they would be easily adapted to other databases containing similar information. Moreover, the characteristics of the source code make it scalable for its application to very large databases.

Building relevant computational models requires series with a good balance between positive and negative compounds. We discussed the importance of avoiding the confusion between absences of observation and negative results. We proposed a practical method that was further illustrated in the modeling application.

One of the key components of the proposed strategy is the definition of finding profiles associated to toxicological endpoints. These associations were a highly valuable contribution of expert toxicologists and allowed transforming collections of observable findings into meaningful *in vivo* endpoints. Its use has been illustrated for 3 selected *in vivo* endpoints. We also have included additional profiles for other endpoints as supplementary information material (Supplementary Table 3). The way in which such assignments were used to infer the properties of the compounds was only a first approach and more sophisticated analyses are being carried out, oriented to improve the accuracy of the assignments.

All in all, we consider that the methods described here represent a significant advance for the exploitation of highly valuable data using computational methods, but we do not claim that they provide a complete solution to such a challenging problem. More work is on its way to address the open problems listed in the "Discussion" section.

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

FUNDING

Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115002 (eTOX), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contributions. Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no. 777365 (eTRANSAFE). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA companies' in-kind contributions.

ACKNOWLEDGMENTS

We would also like to thank Jürgen Funk and Daniel Vettiger from F. Hoffmann-La Roche Ltd. for contributing to the definition of the *in vivo* liver endpoints.

REFERENCES

Agrawal, R., Gupta, A., and Sarawagi, S. (1997) Modeling multidimensional databases. In *Proceedings 13th International Conference on Data Engineering*. IEEE Computer Society Press, pp. 232–243.

Angermueller, C., Pärnamaa, T., Parts, L., and Oliver, S. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878.

Atkinson, F. (2014). Standardiser. Available at: https://github.com/flatkinson/standardiser, last accessed November 20, 2017.

Baader, F., McGinness, D., Nardi, D., and Patel-Schneider, P. F. (2010). *The Description Logic Handbook: Theory, Implementation and Applications* Cambridge University Press, New York, NY, USA.

Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R., and Clementi, S. (1993). Generating optimal linear PLS estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct. Relationships* **12**, 9–20.

Carbonell, P., Lopez, O., Amberg, A., Pastor, M., and Sanz, F. (2017). Hepatotoxicity prediction by systems biology modeling of disturbed metabolic pathways using gene expression data. *ALTEX* **34**, 219–234.

Carrió, P., Pinto, M., Ecker, G., Sanz, F., and Pastor, M. (2014). Applicability domain analysis (ADAN): A robust method for assessing the reliability of drug property predictions. *J. Chem. Inf. Model* **54**, 1500–1511.

Cases, M., Briggs, K., Steger-Hartmann, T., Pognan, F., Marc, P., Kleinöder, T., Schwab, C. H., Pastor, M., Wichard, J., Sanz, F., *et al.* (2014). The eTOX Data-Sharing Project to Advance in Silico Drug-Induced Toxicity Prediction. *Int. J. Mol. Sci.* **15**, 21136–21154.

Codd, E. F. and F., E. (1970). A relational model of data for large shared data banks. *Commun. ACM* **13**, 377–387.

Date, C. J. (1995). An introduction to database systems. *An Introd. Database Syst.* **1**, 839.

Durán, Á., Zamora, I., and Pastor, M. (2009). Suitability of GRIND-based principal properties for the description of molecular similarity and ligand-based virtual screening. *J. Chem. Inf. Model* **49**, 2129–2138.

Free Software Foundation. (2007). GNU GENERAL PUBLIC LICENSE Version 3, 29 June 2007.

Garcia-Molina, H., Ullman, J.D., and Widom, J. (2008) Database systems: the complete book 2nd ed. Pearson Prentice Hall, Upper Saddle River, NJ, USA.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Min. Knowl. Discov.* **1**, 29–53.

GNU General Public License, version 3. (2007) Available at: http://www.gnu.org/licenses/gpl.html, last accessed November 20, 2017.

Hayamizu, T. F., Mangan, M., Corradi, J. P., Kadin, J. A., and Ringwald, M. (2005). The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol.* **6**, R29.

Keenan, C. M., Baker, J., Bradley, A., Goodman, D. G., Harada, T., Herbert, R., Kaufmann, W., Kellner, R., Mahler, B., Meseck, E., *et al.* (2015). International Harmonization of Nomenclature and Diagnostic Criteria (INHAND). *Toxicol. Pathol.* **43**, 730–732.

Kimball, R., Reeves, L., Ross, M., and Thornthwaite, W. (2008) *The Data Warehouse Lifecycle Toolkit Table of Contents* Wiley Pub, New York, NY, USA.

Leffler, J. E., and Grunwald, E. (1963) *Rates and equilibria of organic reactions* Wiley, New York.

López-Massaguer, O. (2017a). eTOX in vivo data extraction tool v1.0. Available at: https://github.com/phi-grib/etox-rdt-extraction-tool, last accessed November 20, 2017.

López-Massaguer, O. (2017b). Scoring tool v1.4.1. Available at: https://github.com/phi-grib/Scoring, last accessed November 20, 2017.

Martin, M. T., Judson, R. S., Reif, D. M., Kavlock, R. J., and Dix, D. J. (2009). Profiling chemicals based on chronic toxicity results from the U.S. EPA toxref database. *Environ. Health Perspect.* **117**, 392–399.

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **3**, 17–31.

Mernik, M., Heering, J., and Sloane, A. M. (2005). When and how to develop domain-specific languages. *ACM Comput. Surv.* **37**, 316–344.

Milletti, F., Storchi, L., Sforna, G., and Cruciani, G. (2007). New and original pKa prediction method using grid molecular interaction fields. *J. Chem. Inf. Model* **47**, 2172–2181.

Molecular Discovery. (2017a). MoKa eTOX version 3.0. Available at: http://www.moldiscovery.com/software/moka/.

Molecular Discovery (2017b) Pentacle 1.0.6. Available at: http://www.moldiscovery.com/software/moka/, last accessed November 20, 2017.

Molecular Networks (2017) AdrianaCode. Available at: https://www.mn-am.com/products/adrianacode, last accessed November 20, 2017.

Mouse Adult Gross Anatomy Ontology. (2017), Available at: https://bioportal.bioontology.org/ontologies/MA, last accessed November 20, 2017.

Mulliner, D., Schmidt, F., Stolte, M., Spirkl, H. P., Czich, A., and Amberg, A. (2016). Computational models for human and animal hepatotoxicity with a global application scope. *Chem. Res. Toxicol.* **29**, 757–767.

Necib C. and Freytag J. (2004). Using Ontologies for Database Query Reformulation. In, *Proceeding on the 18 th conference on Advances in Databases and Information Systems (ADBIS'2004)*

Odersky, M., Altherr, P., Cremet, V., Emir, B., Maneth, S., Micheloud, S., Mihaylov, N., Schinz, M., Stenman, E., Zenger, M. (2004) An Overview of the Scala Programming Language, École Polytechnique Fédérale de Lausanne. Available at: https://infoscience.epfl.ch/record/52656/files/ScalaOverview.pdf, last accessed November 20, 2017.

Pastor, M., Cruciani, G., McLay, I., Pickett, S., and Clementi, S. (2000). GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **43**, 3233–3243.

Pedregosa, F., *et al.* (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830.

PostgreSQL. (2016) Available at: https://www.postgresql.org, 2016, last accessed November 20, 2017.

Ravagli, C., Pognan, F., and Marc, P. (2017). OntoBrowser: A collaborative tool for curation of ontologies by subject matter experts. *Bioinformatics* **33**(1), 148–149.

Sadowski, J., Gasteiger, J., and Klebe, G. (1994). Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Model* **34**, 1000–1008.

Sanz, F., Carrió, P., López, O., Capoferri, L., Kooi, D. P., Vermeulen, N. P. E., Geerke, D. P., Montanari, F., Ecker, G. F., Schwab, C. H., *et al.* (2015). Integrative modeling strategies for predicting drug toxicities at the eTOX project. *Mol. Inform.* **34**, 477–484.

Stonebraker, M., Rowe, L. A., Stonebraker, M., and Rowe, L. A. (1986) The design of POSTGRES. In, *Proceedings of the 1986 ACM SIGMOD international conference on Management of data - SIGMOD '86*. ACM Press, New York, New York, USA, pp. 340–355.

Ullman, J.D. and Widom, J. (2008) A first course in database systems. 3rd ed. Pearson Prentice Hall, Upper Saddle River, NJ, USA.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130.

Wong, L. (2000). Kleisli, a Functional Query System. *J. Funct. Program.* **10**, 19–56.