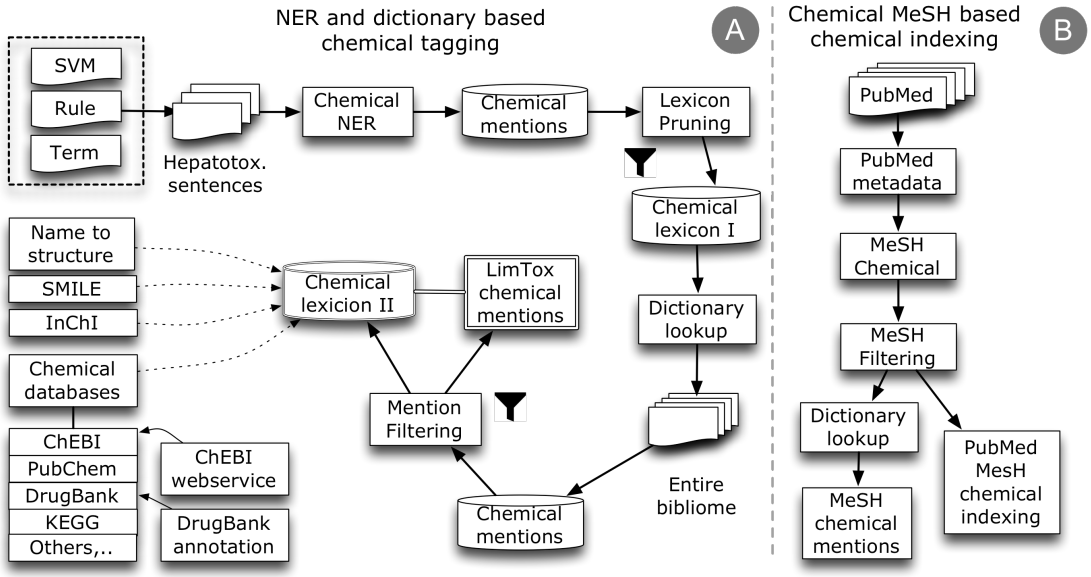
ADDITIONAL FILE

Additional material 1

Recognition of chemical entities and drugs.

A key step for associating chemical compound mentions to toxicity endpoints is the recognition, tagging or indexing of documents with chemicals and drugs. For LimTox, we examined several available resources with the purpose of chemical entity tagging. The examined resources included OSCAR (OSCAR3, OSCAR4)[1], ChemicalTagger[2], the Jochem lexicon (using dictionary name lookup)[3], as well as ChemSpot. We decided to use the ChemSpot tagger (Rocktaschel et al.)[4] for detecting chemical entities because it is able (1) to provide entity grounding to various chemical database identifiers, (2) it can detect systematic, semi-systematic and trivial chemical names, (3) it is able to process effectively large collections of documents and (4) it is freely available/accessible.

Our aim with respect to the chemical tagging process was to primarily focus on compound mentions that are relevant for toxicology studies, rather than using a general-purpose labeling of chemical substance mentions. Additional figure 1 below illustrates the chemical tagging protocols used by LimTox.



Additional figure1. Chemical entity recognition and indexing strategies used by LimTox. Two different approaches were used to associate text to chemical compounds. (A) One approach relied on the ChemSpot chemical tagger in order to generate a compound lexicon, which

[1] http://www-pmr.ch.cam.ac.uk/wiki/Oscar3

[2] http://chemicaltagger.ch.cam.ac.uk/

[3] http://biosemantics.org/index.php/resources/jochem

[4] https://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/chemspot

was then used for text indexing using dictionary-lookup methods (B) The other approach was based on selecting, filtering and look-up of compounds annotated as meta-data (MeSH substance terms) by PubMed.

We carried out dictionary pruning to remove false positive chemical entity mentions. This was done through a two-step dictionary filtering process. LimTox used only the subset of chemical entities detected by the ChemSpot tagger (a hybrid method based on CRFs and chemical dictionaries) that had at least a single occurrence in sentences that were previously recognized as hepatotoxicity-related (see 'Scoring text for Hepatotoxicity' section). We generated from the collection of chemical entity mentions a chemical name gazetteer. The resulting list of chemical names was filtered using a stop word list and filtering rules. The used filtering rules included a list of stop suffixes, stop tokens that should not be present in the last word/token of potential chemical names and removal of quantities, temporal expressions, cell types, DNA codons, and certain common abbreviations corresponding to general English expressions. The top 1,000 names ranked by absolute frequency were manually revised to remove potentially false positive or highly ambiguous names.

Some of the detected chemical names were directly linked to database identifiers or structural information (SMILES and InChI keys) by the ChemSpot tagger. Complementary to this output, association to structural information was also done using name to structure conversion software (name-to-struct version 13.0). We used the ChEBI webservice to retrieve structural information for mentions that were assigned to ChEBI identifiers by ChemSpot.

In addition to the automatic mention tagging of chemical entities, also metadata from PubMed abstracts with chemical MeSH terms were exploited by Limtox (Additional figure 1, subfigure B). Therefore we selected from MeSH metadata the 'Chemical' field corresponding to 'NameOfSubstance' records. We included all cases that had a 'Registry Number' and excluded those that corresponded to EC numbers. We applied also a MeSH term filtering step using a stop word and a stop token list to filter terms containing words indicating that the term corresponded to proteins, agents or systems.

In order to determine if the used chemical entity recognition pipeline was able to detect mentions of chemicals that are relevant for hepatotoxicity, we compiled the list of chemical substances annotated in the ADE-SCAI corpus (Gurulingappa et al. 2012) to be associated to adverse liver events. In this corpus a total of 162 unique compound names were annotated as causing an adverse hepatic reaction. These chemical names were not normalized/linked to any chemical database by the original authors. We performed a manual entity grounding of these compounds to the several databases. All chemical entity mentions could be normalized to at least one database except for 'Lp-TAE', which turned out to be a mixture of various substances and thus was not present in any database. This name was excluded from the dataset. All the other mentions could be linked to a CAS-RN (111 unique compounds). In case of other databases, 91 were present in DrugBank, 102 in MeSH and 88 PubChem compound. It seems that at least for this purpose CAS was the most comprehensive resource. These chemical compounds had associations to a total of 122 PubMed records (354 compound mention-adverse liver event sentences relations in ADE-SCAI). We assessed the recall of LimTox at the level of indexing abstracts either with (a) the unique compound names or (b) with their corresponding database identifiers. A total of 154 out of the 198 chemical name-PMID associations (77.78%) could be detected by the LimTox chemical entity mention recognition approach. The recall when looking at the corresponding chemical database identifiers was slightly better. The recall using CAS-RN was of 79.33% (119 from 150), using DrugBank 81.89% (104

from 127), MeSH 82.14% (115 from 140) and PubChem compounds 83.06% (103 from 124). Most of the missed mentions could have been recovered using additional typographical variants of names already present in the original lexicon.

Additional material 2.

Scoring text for Hepatotoxicity.

Scoring text for hepatotoxicity is not only important in order to detect articles that might be relevant for manual curation of toxicology data, but it is also useful as part of topic-specific retrieval engines and to allow ranking hits when doing keyword or semantic searches. Text data was scored at the level of abstracts and individual sentences. Four different complementary strategies were implemented to allow detection of hepatotoxicity relevant text:

(1) Term strategy (indexing of sentences and abstracts with terms related to adverse hepatobiliary events).

(2) Rule strategy (rule based detection of sentences with co-occurrences of phrases referring to hepatobiliary location/anatomy/cells, i.e. the location trigger and adverse reactions or toxicity events, i.e. the adverse reaction triggers).

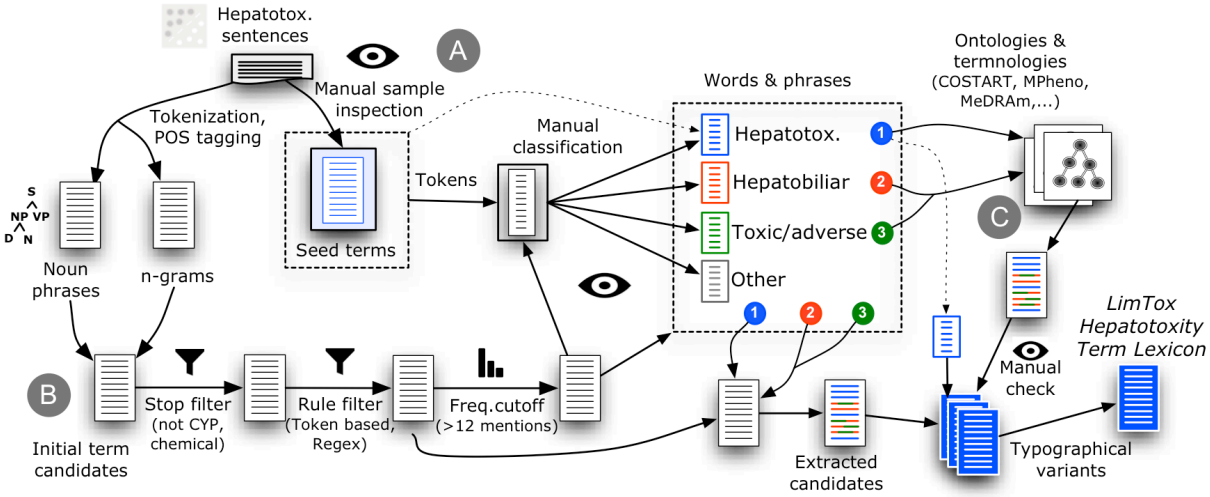
(3) Pattern matching strategy (detection of particular language expressions used to describe chemically induced adverse hepatobiliary reactions).

(4) Supervised machine learning text classifier strategy (machine learning based abstract/sentence text classifiers relying on Support Vector Machines and bag of word text representation models).

Hepatobiliary adverse event term occurrences.

A widely used strategy to associate text to a topic of interest involves indexing documents with terms or phrases that are representative of the topic. The advantage of this method is that it makes human interpretation of results straightforward. An obvious requirement for term indexing is the existence of a suitable lexical resources in the very first place. A specific ontology or thesaurus for hepatotoxicity did not exist, but there were some relevant terms scattered across different ontologies. The strategy used to build a hepatotoxicity lexicon for LimTox is summarized in the flow chart that can be seen in the additional figure 2. First we selected a collection of manually specified seed terms for hepatotoxicity through examination of sentences classified as hepatotoxicity relevant that additionally also mentioned chemical compounds. These sentences corresponded to sentences that were scored as hepatotoxicity relevant by the SVM classifier approach described later in the additional materials section 2. The set of seed terms were then tokenized into words and the corresponding unique words were manually categorized into one of the following semantic classes: (1) hepatotoxicity triggers (e.g. transaminitis, steatosis, hepatotoxic), (2) hepatobiliary related words (e.g. hepatocyte, liver), (3) toxicity/adverse event related words (e.g. toxic, injury, degeneration) and (4) others (not related to any of the previous classes).

.



Additional figure 2. Flowchart of the LimTox hepatotoxicity terminology construction process. There are three main approaches for the LimTox term selection, labelled in the flow chart as (A), (B) and (C). In case of the first strategy (A), a set of manually derived terms were selected based on visual inspection of a sample of sentences scored as hepatotoxicity relevant by the SVM sentence classifier. The second approach (B) is based on automatic extraction of candidate terms directly from the literature by ranking noun phrases and n-grams present in hepatotoxicity sentences. The third approach we used was based on selecting relevant terms from existing ontologies and terminologies.

To perform a more systematic detection of hepatotoxicity relevant terms a semi-automatic extraction, ranking and triage of hepatotoxicity candidate terms was carried out. First noun phrases were automatically extracted from hepatotoxicity sentences using the NLTK toolkit together with MedPost generated POS tags (Smith et al, 2004). The initial set of noun phrases and also n-grams (word bigrams and trigrams) was then filtered using a stop list. This stop list included mainly chemicals, markers and CYPs as well as a set of manually defined filtering rules. Those rules filtered mentions of codons, time and quantity expressions (e.g. mM, min, pH) and numbers. The subsequent list of phrases was ordered based on absolute frequencies within hepatotoxicity sentences. A total of 5,618 phrases were mentioned more than 12 times. These were manually categorized into the previously introduced four semantic classes: 18.69% corresponded to hepatobiliary related phrases, 9.75% to adverse event phrases and 2.47% to hepatotoxicity related phrases. Additionally from those high frequency phrases, word tokens with a particular POS tag (nouns and adjectives) were also classified into these categories.

These manually validated phrases and words were in turn also used to automatically retrieve additional term candidates. Therefore, we applied two selection criteria. The term candidates were added to the LimTox lexicon either if (1) they contained a hepatotoxicity trigger or (2) they contained both a hepatobiliary trigger together with a toxicity/adverse event trigger within the same noun phrase. An illustrative example of the first selection criterion is the phrase '*subacute hepatotoxicity*', while an example of the second criterion is the phrase '*submassive liver necrosis*'. The second term has both a trigger referring to the hepatobiliary system (liver) and a trigger that expresses an adverse event (necrosis).

We used the same term selection criteria to detect hepatotoxicity candidate terms contained in existing ontologies/terminologies. All ontology-derived candidate term were manually validated before adding them to the LimTox lexicon. The following number of adverse hepatobiliary terms were found in existing ontologies/terminologies: Mammalian Phenotype (426), Adverse Events (8), Disease Ontology (576) Gemina symptom (16), Human Phenotype (184), Mouse Pathology (63), COSTART (87), MeDRA (15), CTD database MEDIC lexicon (937), Polysearch lexicon (974), eTOX project toxicology ontology (290). Although we also processed other terminologies, we could only retrieve within them terms related to the hepatobiliar system and not adverse hepatobiliary event terms (e.g. Brenda Tissue (90), Event Ontology (7), vertebrate Homologous Organs Groups ontology (34), Mouse Anatomy (72), Foundational Model of Anatomy ontology (1,262) and Uber anatomy ontology (548)). We needed to apply an extra stop-word post-processing step to exclude particular diseases or adverse liver events that were caused by viral, bacterial or parasitic infections (not due to chemical agents).

The resulting LimTox hepatotoxicity lexicon comprised a total of 29,371 terms (4,141 were manually validated). We enriched this lexicon with automatically generated term variations taking into account rules for plural endings, case variants and hyphenation. The resulting number of term variants was of 200,016.

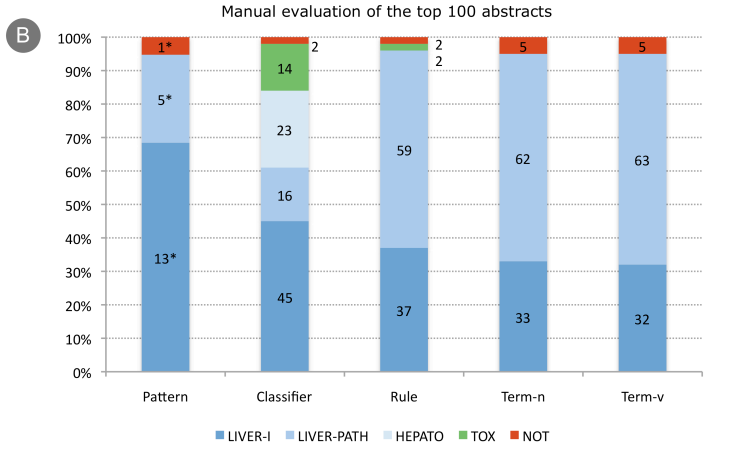
To match the LimTox lexicon terms to text, we used two dictionary look-up settings. The first one was based on matching the automatically generated term variants (Term-v) to text. The second one considered matching of up lower case stemmed terms (normalized) to sentences processed in the same way (Term-n). The retrieval results using these two strategies in Gold Standard sets can be seen in additional figure 3. Documents and sentences were subsequent ranked by the number of detected hepatotoxicty terms. The recall of the term-lookup based method (using term variants) ranged between 87.10 and 97.12.

A

Evaluation of the limtox abstract classification

Method	P top 100	R-CTD	R-MeSH	R-SCAI	R-Manual	R-all	PubMed
Pattern	68.42	56.36	41.61	52.03	66.19	41.67	28,932 (0.21%)
Classifier	45.00	95.86	92.86	89.43	98.08	92.85	767,322 (5.67%)
Rule	37.00	97.69	92.58	100.00	99.28	92.80	541,699 (4.00%)
Term-n	33.00	91.05	85.92	91.06	95.20	85.95	382,627 (2.83%)
Term-v	32.00	93.12	87.10	95.93	97.12	87.16	407,715 (3.01%)
Any	-	99.03	97.25	100.00	99.76	97.31	1082382 (8.00%)
Total	-	1643	12217	123	417	13065	-

B



Additional figure 3. This figure shows the results of the evaluation of the automated detection of hepatobiliary toxicity relevant abstract. The following evaluation datasets were used: (a) the R-CTD set containing articles annotated with the term 'Drug-Induced Liver Diseases' in the CTD database, (b) the R-MeSH set containing records that were indexed by PubMed with the MeSH term 'Drug-induced liver injury', (C) the R-SCAI set, consisting of the subset of DILI relevant abstracts from the ADE corpus (8) and (D) the R-Manual consisting of a small sample set of abstracts that were examined by a last year medicine student and classified as DILI relevant. (A) Evaluation scores for each of the detection strategies, namely the pattern based method, the rule-based system, and the term mention methods (Term-n: mentions of stemmed terms, Term-v: mentions of inflected term variants) or any of the methods, i.e. the abstract was detected by at least one methods. The precision (P) was calculated for the top 100 hits for each method from a random sample of 8000 abstracts based on manual inspection. The recall (R) was estimated using several external datasets as a validation standard (as well as a set of manually labeled abstracts prepared by a last year medicine student, the R-Manual set). The total number of PubMed abstracts retrieved by each method and the corresponding percentages are shown in the last column of this table. (B) Detailed examination of the top 100 abstract evaluations. Each of the top 100 abstracts returned by every method was classified manually into Liver-I (induced adverse hepatobiliary effect), Liver-Path (hepatobiliary disease), Hepato (Hepatobiliary system related), Tox (Toxic effect relevant) and other (not relevant). Note that in case of the hits returned by the pattern-based method for the 8000 sample set only 19 of them contained pattern hits, so in this case the evaluation was based only on these 19 hits.

Rule-based approach.

A general shortcoming of the term based indexing approach is that adverse hepatobiliary events (and also other events) can often be described in the literature a way that does not rely on the use

of a specific term or phrase. To illustrate such descriptive expressions referring to hepatotoxicity consider the following example sentence:

'**Toxic effect** of an antitumor drug *paclitaxel* on morphofunctional characteristics of the **liver** in rats' [PMID:19023985]

Even though this sentence associates a particular chemical to a drug-induced liver injury effect, the authors do not use a specific term built up by consecutive words. Nonetheless, the two elements that are key for hepatotoxicty events (the adverse effect and the actual site or target organ) are co-mentioned in the same sentence.

To recover this kind of expressions, we have constructed a simple rule-based or knowledge-based text processing approach. This strategy explores how information relevant to hepatotoxicity is generally stated in single sentences beyond the use of specific keywords. Through analysis of sample cases we observed that a general property of hepatotoxicity sentences was the existence of expressions referring to the target site (organ, tissue, cell type, molecular entities) affected by the toxic effect together with the description of some adverse, pathologic or toxicity expression.

The rule-based extraction module relied the same principle as previously described for selecting automatically candidate hepatotoxicity terms. This implies that it required that both a location trigger term and an adverse event trigger term had to co-occur together in a particular context (sentence). The rule-based method used a total of 852 manually defined hepatotoxicity, 960 hepatobiliary and 552 toxicity trigger terms. These were part of the trigger lists compiled for the term selection method (although some triggers that were too ambiguous were finally excluded). The main difference between the rule-based system and the term-selection process was that the contextual window used by the rule-based approach consisted of entire sentence instead of noun phrase or word n-grams. We used a heuristic sentence scoring scheme to weight the output of the rule-based system. In short, this scoring mechanism took into account the total number of co-occurrences between hepatobiliary terms and adverse effect terms in a sentence, the respective relative sequential order within the sentence and the relative distance measured by the number of word tokens between them. The relative position and distance features were considered only for the closest co-occurrences between adverse and hepatobiliar triggers (measured in word tokens). The rule-based scoring of entire abstracts consisted of taking the sum of the corresponding sentence scores. This method had a high recall, as shown in the results of additional figure 3. The recall of this method ranged between 92.58 and 100%.

Pattern-based approach

Under certain circumstances users favor high recall results. For instance in case of literature curation, high recall triage of articles that will be later manually curated can be of importance. Under settings where only limited human workload is possible, or when text mining results are intended to be directly used to populate a knowledgebase, high precision results are often desirable. Pattern-based methods constitute one approach that is still widely used in information extraction to achieve high precision results. Two kinds of alternatives when looking at pattern-based techniques are purely statistical pattern learning on the one side and hand-crafted patterns on the other.

We have explored the use of hand-crafted patterns for detecting drug-induced liver injuries (DILIs). These patterns were constructed through manual examination of sample instances as well as by using statistical, grammatical and gazetteer-matching selection criteria of candidate templates. The later was done in order to improve selection of frequent expressions suggesting DILIs.

As an initial constraint, the pattern-based approach required co-occurrences of two semantic types, namely the *agent* (chemical compound or drug) and the *target* site (hepatobiliary system) within sentences.

A preliminary set of manual text patterns were constructed from sample sentences with agent-site co-occurrences, which also had a high hepatotoxicity sentence classifier score. Text patterns corresponded to the minimal text spans that referred to a causal relation between a chemical and an adverse hepatobiliary event.

Chemical mentions as well as the adverse event terms were masked with a semantic class label. This allowed generalizing text pattern by transforming them into a sort of template. Consider the following illustrative example pattern/text template:

Pattern: <***agent:chemical***> *in inducing* <***target***:**adverse**> ➔ [troglitazone in inducing hepatotoxicity]

Here 'in inducing' is the core of the text pattern, describing a causal adverse effect relation pattern and containing a causal relation trigger verb (induce). The collection of used extraction pattern can be downloaded from the LimTox resources webpage[5]. Additionally we examined another source of sentences describing chemically induced adverse reactions, namely the ADE-SCAI corpus. First, all sentences of this corpus were divided into (a) those that described adverse hepatobiliary effects and (b) other types of adverse reactions. This second collection of 6,466 sentences was inspected manually to construct handcrafted text patterns. The collection of DILI related sentences from the ADE-SCAI corpus was held back as a validation set for the relation extraction task. Adverse events and the chemical entities were masked to generate templates. We kept only textual patterns that were not biased towards a particular subtype of adverse effect. This was done to make sure that they would be able to recover any causal adverse event relation. SCAI corpus derived example pattern:

Pattern: <***target:adverse***>: *caused by* <***agent***:**chemical**> ➔

[*ADVERSE caused by CHEMICAL* in patients with inflammatory bowel disease [PMID:10203437]

One obvious drawback of the previous two selection criteria has to do with the fact that a large number of surface grammatical structures can in practice refer to adverse events in the literature.

In order to prioritize those that are more commonly used in the literature, we considered statistical (raw frequency), grammatical (POS labels of words) and gazetteer (a list of 21 trigger terms) information. From sentences that contained co-occurrences of chemicals and adverse hepatobiliary terms, the text fragments joining the co-occurrences within the sentence were selected (stripping off the left and right flanking words. In case of the previous example this would be 'caused by'. These inner connection text fragments were tokenized and only those fragments retained that had

[5] https://github.com/inab/etox/tree/master/lexicon

between 1 and 9 words. The top 3,000 patterns, ranked based on absolute frequency, were manually revised. The remaining list of patterns were processed based on whether: (a) they were less than 5 tokens long, (b) they contained at least one verb or noun (excluding auxiliary verbs and verbs with less then 4 characters in length) and (c) they had a maximum length of 6 words together with mentions of at least one of 21 predefined relation trigger terms.

To cover scenarios not strictly limited to the inner connection text fragments, one complementary selection criteria of candidate patterns was considered. This case entailed the selection of up to 3 words on the left of the first semantic class label mention together with corresponding inner connection fragments of length 1-3 words. Within the left flanking segment a nominalized form of the trigger term had to be mentioned. Moreover, only those segments were chosen that did not start with certain POS tags ('IN','CC','TO','DT',':','CD','RB','PRP$','PRP',',','-NONE-','"``","WRB','WP','WDT','RBS','RBR','POS','MD'). Nominalizations are quite common and important in the biomedical literature and they are also associated with some alterations in terms of syntactic constructs and consequently word order (Cohen et al, 2008).

The top 6,000 patterns ranked by frequency that fulfilled the previously describe selection criteria were then manually validated.

Joining all the manually validated patterns, a total of 2,926 pattern templates were obtained, corresponding to 2,896 case insensitive patterns with masked hyphens. These patterns were used as templates to recognize matching phrases within sentences. For abstracts ranking purposed we used the total number of corresponding sentences that contained pattern matches. The resulting pattern-based approach was rather competitive in terms of precision was can be seen when looking at the additional figure 3. This approach was more reliable in detecting chemically induced adverse hepatobiliary effects. Nonetheless, there was also a considerable drop in recall when compared to the other approaches.

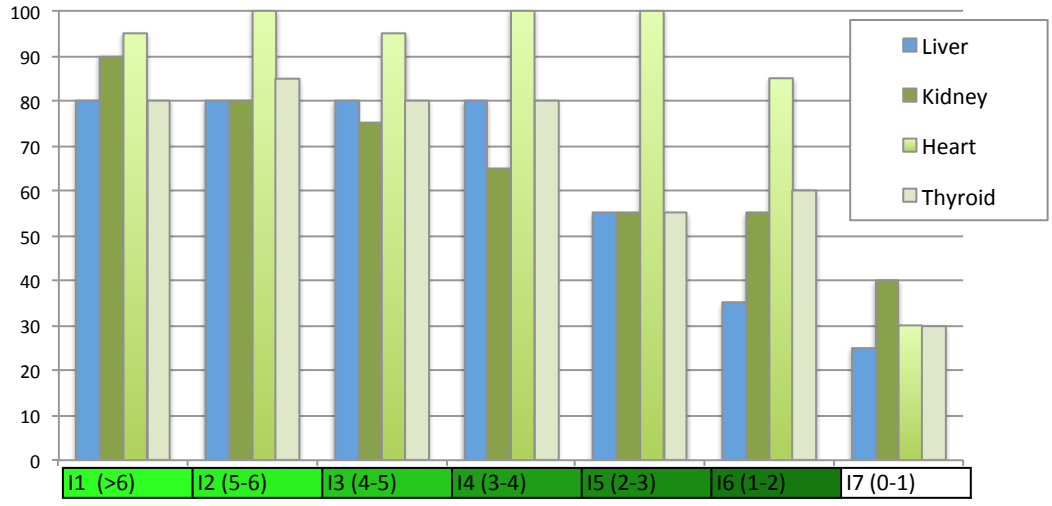
Text classifier strategy

The hepatotoxicity abstract classifier was trained on a balanced set of 10,984 abstracts. The positive training data contained records relevant for drug-induced liver damage selected by either the keyword or rule based approaches previously introduced. The negative training set was a random sample of PubMed abstracts of the same size. To determine the quality of the used training data a sample of 100 abstracts was manually inspected, 83% corresponded to DILI relevant documents and another 12% to adverse liver events caused by alterations in genes and gene products (mainly at the level of gene expression). The abstract classifier consisted of a linear kernel SVM classifier that used a bag-of-words (BOW) representation model of the text and unigram term frequency as feature weights. The aim of this strategy was to implement a high recall system that enables classification and ranking of hepatotoxicity relevant articles. The resulting classifier model was applied to score the entire set of abstracts contained in the PubMed database. 5.67% of all abstracts were scored as relevant for adverse hepatobiliary events. For evaluation purposes, and in order to estimate whether the classifier was able to detect records annotated as being relevant for hepatotoxicity, we used several independent evaluation data sets. The obtained recall results can be seen in additional figure 3. The resulting classifier model was able to recover between 89.43% and 98.08% of the records annotated as DILI relevant from various datasets. The following evaluation datasets were used: (a) the R-CTD set contained articles annotated with the term 'Drug-Induced Liver Diseases' in the CTD

database, (b) the R-MeSH set contain records that were indexed by PubMed with the MeSH term 'Drug-induced liver injury', (C) the R-SCAI set, consisting of the subset of DILI relevant abstracts from the *ADE* (adverse drug effect) corpus[6] from the Fraunhofer Institute for Algorithms and Scientific Computing (*SCAI*) (Gurulingappa et al. 2012) and (D) the R-Manual consisted of a small sample set of abstracts that were examined by a last year medicine student and classified as DILI relevant.

Additionally to the abstract classifier also two distinct hepatotoxicity sentence classifiers were constructed. The balanced training set for both consisted of 28,203 sentences. The positive training set was selected through a term and rule based selection process while the negative training sentences were randomly chosen from PubMed. To determine the quality of the positive training sentences a sample of 100 was manually inspected, 85 % corresponded to DILI relevant sentences and 15 % to adverse liver events (only 1% was neither of both). The classifier relied on word n-gram features with a range of 1-4 lowercase tokens and term frequency-inverse document frequency term weighting. For this purpose we exploited classipy, a command-line tool originally developed in our lab that can be used to develop advanced text classifiers using SciKit-Learn. The same pipeline was used to generate sentence classifier models for additional toxicity endpoints, namely nephrotoxicity, cardiotoxicity, thyrotoxicity and phospholipidosis.

To estimate the precision of the sentence classifier scores, we have evaluated randomly selected sentences for the same score intervals as highlighted by the LimTox sentence scoring color schema. Randomly selected sentences were chosen for each of the toxicity endpoints restricted to score intervals. Those sentences were manually examined whether they described adverse events for each of the organ systems examined. The obtained precision results for each of the endpoints and score intervals are shown in the additional figure 4, together with the color schema used for the LimTox application.



Additional figure 4. This figure shows manual evaluation of precision of the sentence classifier using randomly selected sample sentences for each of the score intervals, ranking from SVM classifier scores above 6 (interval 1, I1) to 0 (interval 7, scores between 0 and 1). The bars of the figure correspond to the percentage of sentences that were classified as related to adverse events for each of the corresponding organs (liver, kidney, heart and thyroid).

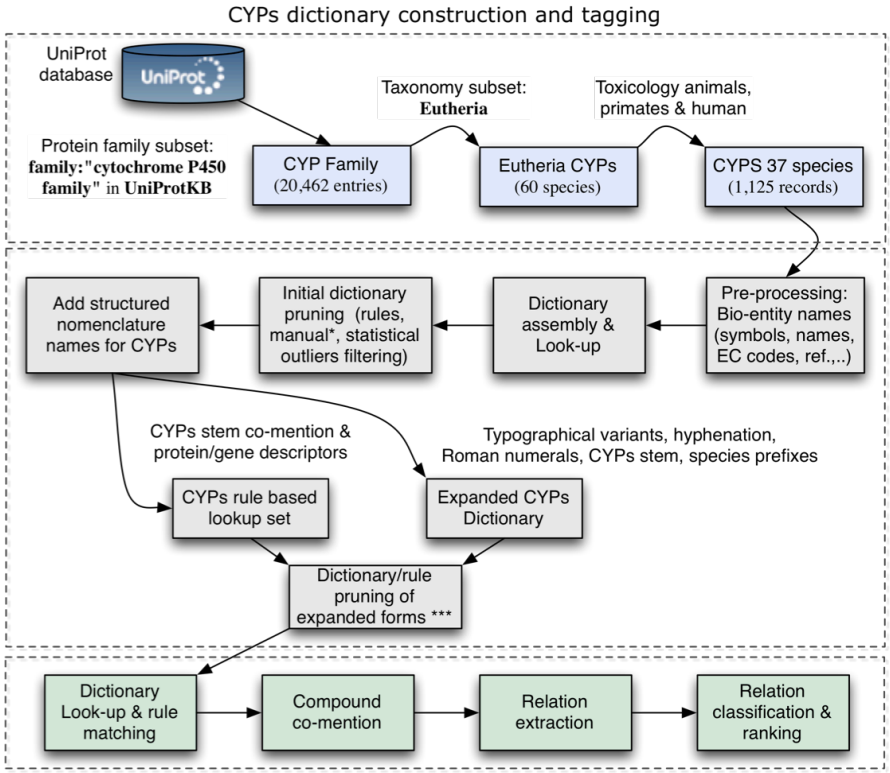
6 https://sites.google.com/site/adecorpus

Additional material 3

Extraction of CYPs relations to chemicals

Due to the fundamental function of cytochromes P-450 (CYPs) in the xenobiotic metabolism of drugs, they are key for understanding toxic effects related to the metabolism of compounds. CYPs play an important role in chemical biotransformation reactions that can result in activation of chemical compounds into toxic species, or detoxification/enhanced elimination of drugs from the organism. Moreover, they have been associated to the metabolic activation of pre-carcinogens and are therefore an interesting target for characterizing predisposition to certain cancer types (Rodriguez-Antona and Ingelman-Sundberg, 2006). Over 57 active human P450 genes have been described so far. They correspond mostly to polymorphic genes that result in different isozymes, with characteristic substrate specificities. We implemented a pipeline for extracting automatically relevant CYPs mentions and interactions from PubMed abstracts, not only for human CYPs but also for CYPS from animal species that are relevant for toxicology studies. Additionally also relations from full text articles and agency reports were extracted.

The initial detection of CYP mentions from the literature is a crucial step for subsequent relation extraction approaches. We addressed the recognition of CYPs in text using a combined strategy relying on a multi-species gene lexicon, semi-automatic lexicon enrichment/pruning and a rule-based approach. Additional figure 5 shows a schematic flowchart describing the used CYPs mention detection approach.



Additional figure 5. Flowchart illustrating the process of CYPs mention recognition, covering from the CYPs lexicon construction and also the rule based CYPs recognition.

An initial CYPs gene/protein name and symbol dictionary was derived from the UniProt database. Therefore we carried out a search in UniProt using as a query: family:"cytochrome P450 family" to retrieve all members of this protein family, resulting in a set of 20,462 entries. These hits covered CYPs contained in the UniProt database, including a total of 287 human hits (60 reviewed and 227 un-reviewed cases). To focus on those species that are interesting for toxicological studies, we selected a taxonomic subset corresponding to Eutheria (60 species), which was then manually filtered, resulting in 37 species (with a total of 1,125 associated UniProt CYPs entries). This seed lexicon was manually examined to remove cases of highly ambiguous names/symbols and non-informative terms (e.g. 'Uncharacterized protein' or 'Putative uncharacterized protein'). For each record, a name following the systematic nomenclature for the various isozymes was added wherever possible. The CYP nomenclature takes into account protein/gene sequences to group them into families and subfamilies. According to the current hierarchical nomenclature conventions, individual CYP names are supposed to follow certain rules, namely: They should start with the common root or prefix (CYP), which are then followed by a number corresponding to the gene family (e.g. CYP1). Thereafter a letter standing for the CYPs subfamily is added (e.g. CYP1A), followed by a number that characterizes the gene (polypeptide) (e.g. CYP1A1) (Cupp and Tracy, 1998).

From the 1,878 unique names contained in this 'baseline' CYPs lexicon, only 269 followed the nomenclature conventions. Less than a third of the unique baseline names (i.e. 537) could be identified in PubMed sentences, although the total number of mention was of 87,364. To increase recall, we applied semi-automatic expansion of the CYP names using manual rules to account for typographical variations (alternative use of hyphenation and spaces, Roman and Arabic numeric expressions, upper case, lower case and capitalized versions of names). Through manual inspection of a collection of randomly selected abstracts known to be relevant for CYPs (cited in UniProt records), we defined a set of rules for generating variants from the official nomenclature names. These rules took into account combinations based on various root forms commonly used for CYPs[7], alternative upper and lower case forms of the CYP subfamily letters and both Arabic and Roman numbers for CYP families. Also species specific prefixes for some of the organisms were added (e.g. 'm 'for mouse and 'Rn' for rat). This resulted in an expanded lexicon of 243,657 unique CYPs names. Some highly ambiguous names were deleted[8]. The CYP nomenclature guidelines were encoded into a pattern matching script that identified potential mentions of cytochrome P450 enzymes. The rule based mention detection approach was able to detect non-continuous text strings referring to a particular CYP, by exploiting the hierarchical nomenclature properties. In brief, it required the mention (within a sentences) of a number followed by a capital letter and again a number (with and without spaces and hyphens) together with one of the following trigger tokens: CYP, Cytochrome or P450. The rule based method was able to cope with enumerations or lists of CYP mentions, where the actual CYP name consists of a non-continuous string of text, as is the case in the following example sentence: 'The effect of obesity on the cytochrome P450 1 A2, 2C9, 2C19, and 2D6 isozymes is

[7] Root form: CYP, Cyp, P450, P-450, P450 (CYP), Cytochrome-P450, etc.

[8] Filtered names: P52, P24, LDM, TXS.

inconclusive'. In this example sentence four different CYPs are mentioned but only the first one, namely 'P450 1A2' would have been detected by the dictionary-based approach.

A total of 250,740 CYPs sentence mentions were detected in PubMed abstracts, 218,803 were recognized by the dictionary look-up method and 31,936 by the rule-based system. For these mentions, the co-occurrences with automatically tagged chemical compounds were generated. A total of 242,870 chemical compound-CYP co-occurrences were detected in PubMed, corresponding to 23,209 unique compound names and 1,940 unique CYP names extracted from 92,327 sentences (39,779 abstracts). Although co-occurrences can be useful to provide general statistical associations between entities, they are not sufficient to label the actual type of relation existing between entities (given the context of mention). Three types of drug-CYPs relation are of particular practical importance for pharmacology and toxicology, namely induction, inhibition and metabolism relations. The induction relation type covers relationships, where a chemical causes an increase in expression or activity of a particular CYP. The inhibition relation type refers to the relation between chemicals that cause a decrease in the expression of a particular CYP gene product or bind to a CYP and inactivate it. Finally, metabolic relations in this context refer to relations between a CYP and a chemical that is biotransformed by it (substrate) or that is the result of such a metabolic reaction (product), including also intermediate compounds generated during the transformation process. The relation extraction strategy used here was similar to the chemical-term relation extraction approach; in the sense that both a pattern/rule based method together with a machine learning sentence-based relation classifier were used.

The rule based approach relied on a list of relation trigger terms that were compiled for each of the three relation classes. The relation triggers were generated by manual revision of POS-tagged verbs associating CYPs and chemicals from the co-occurrence sentences. Verbs were ranked based on their absolute frequency. The top ranking verbs were inspected and classified according to their relevance for these three relation classes. Additionally also synonyms and triggers defined *ad hoc* were included. This resulted in a set of 119 induction triggers terms, 128 inhibition trigger terms and 407 metabolism trigger terms. In order to generate pattern-matching rules for filling template slots of relevance for each relation type the connecting text fragments between the mentions of the CYPs and chemicals were extracted. Only those fragments formed by less than 6 words and also mentioning at least one trigger were retained. The frequency ranked list was then manually inspected to derive relation extraction patterns. Those patterns were classified into the three relation type categories. This resulted in 973 induction patterns, 1,092 inhibition patterns and 1,851 metabolism extraction patterns.

Example CYPS relation patterns are:

(1) induction: '[CHEMICAL] *induced the expression of* [CYPs]'

(2) inhibition: '[CHEMICAL] *is a strong inhibitor of* [CYPs]

(3) metabolism: '[CYPs] *the enzyme that converts* [CHEMICAL['

Those patterns were then used to mine the entire set of CYPs-chemical co-occurrence sentences, detecting 3,192 CYPs-chemical-sentence triplets for induction, 5,159 for inhibition and 4,833 for metabolism. When looking only at the actual interactors, that is the CYPs and chemical mentions regardless redundancy in terms of multiple sentences providing the same relation evidence, a total

of 1,712 unique induction, 1,933 inhibition and 2,679 metabolism chemical-CYPs relations could be extracted by this technique.

To overcome potential recall limitations of pattern-based methods, know to have difficulties in handling long-range associations between entities (mentioned far apart within the same sentence), also three SVM relation sentence classifiers were implemented, one for each relation type. As balanced training set for the classifiers, sentences detected by the pattern-based technique were used as positive training data and randomly selected chemical-CYPs co-occurrence sentences were chosen as negative training data. As features, n-grams (n=1,4) were used, previously masking mentions of chemicals and CYPs. The results using 5-fold cross validation of the classifiers were: precision 92.8%, recall 89.8% and f1-score 91.3% for the induction relations; precision 91.3%, recall 91.5% and f1-score 91.4% for the inhibition relations and precision 88.1%, recall 88.5% and f1-score 88.3% for the metabolism relations.

Each of the classifiers was used to score the entire set of CYPs-chemical co-occurrence sentences. The induction classifier returned a total of 23,412 chemical-CYP-sentence triplets, while the inhibition classifier retrieved 22,354 and the metabolism classifier 26,359 triplets.

When comparing the results of the pattern-based methods to the SVM classifiers, 2,897 of the triplets were detected by both methods in case of induction, 4,833 in case of inhibition and 4,024 for metabolism. This means that 90.76% of the pattern results were also confirmed by the induction relation classifier, 88.64% by the inhibition classifier and 83.26% by the metabolism classifier.

In order to have a better picture of the precision of these relation extraction methods, random samples of 100 relations detected by each method were selected and then manually examined. The resulting precision of the pattern-based induction extraction was of 96%. When looking at the false positive relations, two of them corresponded to NER errors (e.g. PCOS was detected wrongly as a compound in case of: 'PCOS theca cells'). The other two FPs corresponded to other relationships not being induction. The precision of the pattern-based inhibition relation extraction was slightly better (98%). One of the FPs could be attributed to a NER error. In case of the metabolism relation, the pattern technique obtained a slightly worse precision of 95%.

When looking at the predictions of the SVM relation classifier, the precision was considerable lower, in case of the induction relation it was of 46%. The number of errors for this relation type was mainly due to incorrect NER results, which were also considerably higher (11% of the total induction relations examined). Most of the NER errors corresponded to acronyms of cell lines and some also to gene symbols instead of chemicals. Many of the wrongly extracted relations were due to the presence of multiple, complex relations described in the same sentence.

The obtained precision of the inhibition relation classifier was of 54%. In case of the FP relations, many of them were between the compound *tamoxifen* and the CYP *aromatase*. Aromatase inhibitors are often described as an alternative treatment to tamoxifen in the literature. Most of the other errors were due to mentions of multiple CYP-chemical relations in the same sentence resulting in an additional level of ambiguity. The system in those cases often returned the incorrect association pairs.

The metabolism relation classifier had a precision of 69%. The metabolic relations were unexpectedly easier for the machine learning method. Also in this case, most of the FPs were due to multiple CYP-chemical relations described in a sentence, corresponding only a fraction of them to metabolism relations. One strategy to account for this issue would be to propose a weighting scheme based on the number of relation pairs mentioned in a sentence (to down-weight cases were many chemicals and CYPs appear).

Another important aspect when evaluating the performance of the LimTox CYPs relation extraction system was recall. The recall was estimated by comparing the extracted relations to the annotations from two databases: SuperTarget (Günther et al., 2008) and SuperCYP (Preissner et al., 2010). Both of these databases contain relations between CYPs and chemicals covering inhibitor, inducer and substrate interactions. Annotations from both collections were harmonized to CAS-RN (chemicals) and canonical CYP names and UniProt accession numbers (CYPs). The evaluation was done for the subset of human CYPs. The joined Gold Standard set contained 457 induction, 1,396 inhibition and 1,836 substrate relations. Additional table 1 provides not only an overview of the previously described CYPs results but also the recall evaluation of the CYPs relation extraction methods compared to a baseline defined by sentence co-occurrence. This baseline recall ranged between 60 and 66 percent for the various relation types, showing that a considerable number of relations were not detected at the level of co-mention in single sentences. Examining the actual results revealed that many of the Gold Standard compounds were not detected at all in any CYPs sentences. When comparing the recall results of the pattern or SVM extraction methods to the baseline co-occurrences showed that these two methods were relatively competitive for cases were entities indeed are co-mentioned in text. Overall, a slightly better recall was obtained for the substrate relations, while induction and inhibition relations had very similar recall numbers.

Data	Induction	Inhibition	Metabolism*
Relation trigger term	119	128	407
Relation patterns	973	1,092	1,851
Triplets (pattern)	3,192	5,159	4,833
Pairs (pattern)	1,712	1,933	2,679
SVM cross-validation (F-score)	91.3%	91.4%	88.3%
Triplets (SVM)	23,412	22,354	26,359
Precision sample (pattern)	96%	98%	95%
Precision sample (SVM)	46%	54%	69%
Recall co-occurrence (all)	60.61 %	66.26 %	64.77 %
Recall pattern (PubMed)	39.39%	39.26%	43.25%
Recall pattern (all)	42.23%	41.55%	46.02%
Recall SVM (PubMed)	42.89%	41.47%	47.00%
Total Gold Standard	457	1,396	1,836

Additional table 1. Overview of the CYPs text mining results in LimTox. Triplets: chemical-CYPs-sentence; Pair: chemical-CYPs. *In case of the recall evaluation metabolic relations examined consisted only of substrate relationships.

Extraction of liver marker alterations

During the clinical examination of patients, a widespread strategy to detect hepatocellular injuries and cholestasis (blockage of bile flow from the liver) relies on measurements of serum liver enzyme

activities, sometimes called liver enzyme tests or liver function tests. In case of hepatocellular injuries, increased activities of certain enzymes within hepatocytes are frequently detected. Therefore we included in the LimTox system the automatic extraction of relationships between chemicals and the most commonly studied entities measured in biochemical liver assays. A total of 17 liver markers (13 proteins, 3 chemicals and 1 generic term) were carefully selected by reading toxicology review studies and relevant sections of an introductory toxicology book. Additional table 2 provides an overview of the used markers together with some additional overview information. In the literature, and especially in short abstracts, authors often do not specify the particular liver marker measured, but refer to it using a generic term. An entity type for such generic mentions was also included (e.g. liver tests, liver function test, liver transaminases, aminotransferases). The marker lexicon was derived from databases (UniProt and ChEBI) and enriched manually by examining the results returned by the Acromine system for the marker acronyms[9], resulting in a lexicon of 1,590 marker names. The recognition of liver markers was addressed using a dictionary look-up approach together with an acronym disambiguation-filtering step.

Identifier	Name	Short	Aliases	Mentions	PubMed	Pattern Up	Pattern Down
CID 10964	malonyldialdehyde	MDA	127	58,159	56,336	1,124	425
CID 124886	glutathione	GSH	27	223,342	196,603	3,476	28,629
CID 5280352	bilirubin	-	4	46,604	42,615	1,113	109
P00367	glutamate dehydrogenase	GDH	45	11,816	11,129	44	73
P00390	glutathion reductase	GRx	52	11,667	10,796	97	151
P00441	superoxid dismutase	SOD	79	117,409	109,205	1,150	745
P04040	catalase	CAT	20	83,906	72,242	1,047	972
P07195	lactate dehydrogenase	LDH	118	71,005	65,388	2,933	500
P07203	glutathion peroxidase	GPx	58	35,242	32,743	367	366
P10696	alkaline phophatase	ALP	49	94,967	88,122	2,889	589
P17174	serum glutamate oxalate transaminase	SGOT	252	45,263	40,351	1,448	119
P19440	gama-glutamyl transferase	GGT	217	30,210	28,545	522	231
P24298	serum glutamic pyruvic transaminase	SGPT	346	61,809	54,281	2,783	134
P28838	leucine aminopeptidase	LAP	90	6,118	5,353	21	12
Q00796	sorbitol dehydrogenase	SDH	52	4,088	3,379	37	40
Q9H0P0	5'-nucleotidase	5'-NT	43	6,876	6,564	73	76
Unspeficic	liver tests	-	11	35,110	29,698	1,740	44

Additional table 2. Overview of the liver marker text mining results. Aliases: number of synonyms and variants in the marker dictionary for that particular marker; Mentions: number of mentions in the entire document collection; PubMed: number of PubMed abstract mentions; Pattern up: sentence triplets (marker-chemical-sentence) detected with the pattern approach for marker increase; Pattern down: the same as the previous number but for marker decrease. SVM up and SVM down correspond to the number of triplets detected by the SVM relation classifier for increase and decrease. Chem.: number of unique chemical names extracted with the pattern relation approach (for both increase and decrease); CAS-RN: total number of CAS-RNs that were detected with the pattern marker relation method. Chem. SVM and CAS-RN SVM correspond to the same type of results but for the SVM relation classifier method.

A simple rule based system to determine if there is an increase or decrease of liver markers following drug administration was also implemented. This system relied on a list of trigger terms for the different relation types. The increase (up) relation patterns were based on 167 manually defined

[9] www.nactem.ac.uk/software/acromine

trigger terms and 1,925 patterns[10], while the decrease (down) relation patterns were based on 85 trigger terms and 336 patterns[11]. Compared to the other pattern-based relation extraction approaches introduced earlier, in case of the marker relations, there were some pattern templates that contained both a slot for the marker entity and the chemical entity (e.g. '*CHEMICAL increases MARKER level*'), while other patterns did not require the chemical entity as part of the pattern itself, but rather the co-occurrence anywhere in the sentence (e.g. '*MARKER is > uln'*). The number of extracted relation triplets for each marker and the associated chemicals detected through the pattern-based approach are shown in additional table 2. For most of the markers it was more frequent to find an increase relation triplet rather then a reduction. For instance in case of SGPT a total of 2,783 increase triplets were extracted, while only 134 reduction triplets could be obtained. One outlier was the marker *glutathione* with very high number of *decrease* relations.

Sentences detected by the pattern-based approach were afterwards used as positive training set to construct SVM relation classifiers for each of the two relation types. As negative training data a randomly selected set of marker-chemical co-occurrence sentences of the same size was used. In case of the increase relations the balanced training set comprised 35,416 sentences and in case of the decrease relations it consisted of 24,139 sentences. The 5-fold cross validation result for the increase relation classifier was: 94.2% precision, 91.6% recall and 92.9% F-score. In case of the decrease relation classifier the precision was of 95.5 with a recall of 92.9% and an F-score of 94.2%.

[10] Example increase trigger terms: increase, increment, elevate, two-fold, upper limit of normal. Example increase patterns: MARKER elevations, MARKER is $>$ uln.

[11] Example decrease trigger terms: drop, reduction, sinking. Example decrease patterns: decrease MARKER level, low total MARKER.