

Early Identification of Potential Drug Safety Issues from Diverse Literature Resources

Sarah J McQuay

Linguamatics Ltd, St John's Innovation Centre, Cowley Road, Cambridge, CB4 0WS, UK

Telephone: +44 1223 421360 Email: sarah.mcquay@linguamatics.com Website: www.linguamatics.com



Drug toxicity remains one of the major reasons why more than 9 in 10 new drug candidates which enter clinical trials fail to reach the market. However, relevant information that could be useful for predicting the safety of novel drugs may already be held within pharmaceutical companies' corporate archives, reported in the form of text and/or semi-structured data tables. In addition, there are vast amounts of information in the public domain concerned with pharmacological interactions, clinical trial information and news about drug withdrawals that might provide relevant insights. We show here a method of text mining these diverse document resources to link information, uncover hidden knowledge and help predict the toxicity profile of potential blockbusters.

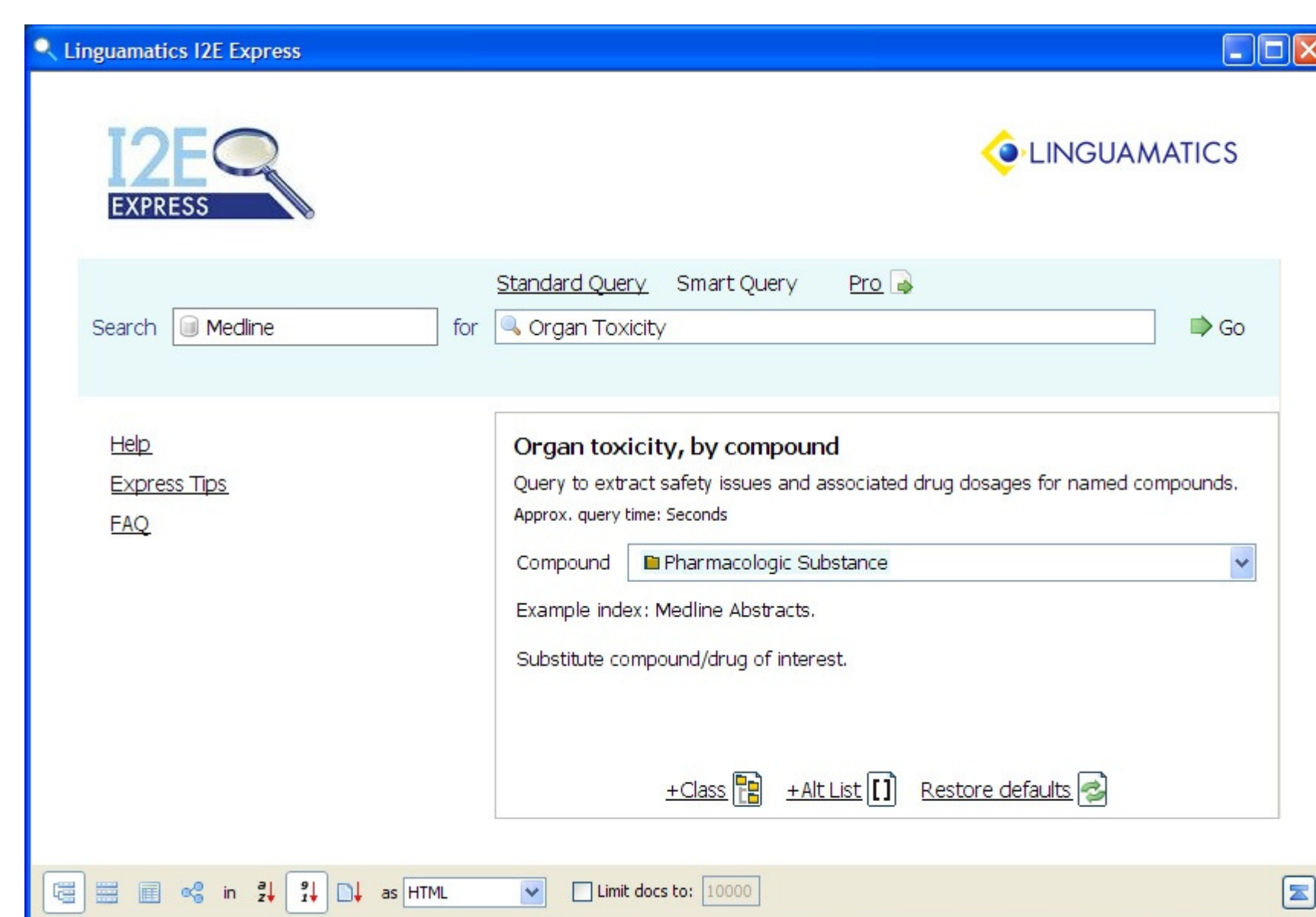


Figure 1. Smart query entitled "Organ toxicity, by compound" shown in the I2E Express interface. In the query box labelled "Compound" the user can substitute a keyword, list of terms (+Alt List) or one or more concepts from a thesaurus/taxonomy (+Class), then click "Go".

Introduction

Linguamatics I2E is a semantic knowledge-discovery platform. It provides an interactive experience similar to web search, but with the extra ability to output high quality facts and relationships (Milward *et al.*, 2005). The system is highly scalable, enabling search over the whole of Medline, large sets of full text documents, patents or an archive of internal reports. I2E can be optimized for a specific domain using relevant terminologies of concepts and synonyms.

Searches may range from simple co-occurrence within a document or a sentence, to very precise linguistically structured patterns (Milward *et al.*, 2006). Queries can include broad semantic classes such as any human gene or any disease, general classes like P450s or cancer, or a specific concept like retina or phototoxicity.

Queries can be saved, shared, reused or combined, for example providing a systematic profile of a compound, collated from various documents. An expert's search strategy can be presented as a smart query, cloaking the sophisticated parts and exposing just one or two items for a user to specify themselves, such as a lab code or therapeutic area, in an annotated form-filling style template.

I2E output is presented as a structured table of results in a choice of formats. These include web pages, XML files, formats suitable for export to databases, Microsoft Excel® spreadsheets, mindmaps and network graphs, allowing the visualization of direct and indirect connections (Milward & Milligan, 2007).

I2E's flexibility is particularly useful for text data mining where it is beneficial to adopt different approaches for different requirements. For example, a sentence co-occurrence strategy may be best to retrieve all mentions of a compound with a certain target, to achieve maximum recall. However, searching for precise linguistic patterns would be necessary to pinpoint a drug-event relationship with required precision.

Method

In this example, a smart query (see **Figure 1**) has been developed to mine available text for drug safety indicators. This query allows the end-user to specify a particular drug or a class of compounds using the built-in class chooser (see **Figure 2**), by typing a set of keywords into the form or uploading a plain text file which has one search term per line.

This single smart query is the doorway to a series of individual queries, which seamlessly run and combine during the search execution, to provide a profile of information in a single table of results. A power user or experienced information professional may wish to explore, and if necessary modify or augment, the underlying search strategy. In this way, they can customize the overall query to extract information that is tailored to their own specific requirements.

The underlying queries in this example smart query contain precise search terms used to identify key information. For example, to find information about adverse events in the liver there are two sets of terms: one set looks for liver-specific terms; the other looks for toxicity specific terms. These two parts of the search can both return the same document hit(s), for example, "hepatotoxicity" will match both terms. The term "hepatotoxicity" will be matched against two wildcard matches: "hepato*" (one of the liver-related terms; others included "liver" and "hepatic") and "*toxicity" (one of the adverse event related terms; others included "side-effects", "adverse events", "sensitivity").

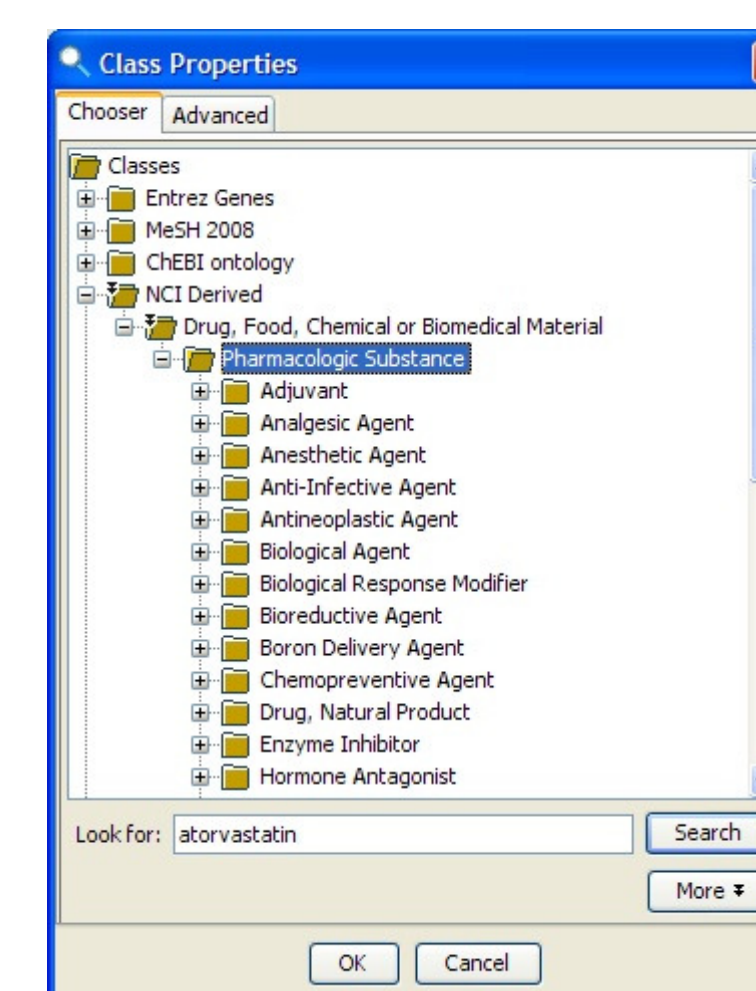


Figure 2. Class Properties dialog box.

Finding a compound to substitute into a smart query can be as easy as searching or browsing for a specific name in a tree of available terms. Here, atorvastatin is shown in the 'Look for' box. Selecting a folder will use the entire class, for example Pharmacologic Substance is shown highlighted here.

Results

Figure 3 shows the typical output from this smart query as a web page showing a cluster table containing evidence linking adverse events to compounds at particular dosages. It is also possible to cluster the results by compound, dosage and tissues of interest to provide a rapid overview prior to more exhaustive analysis.

Clickable links are provided to highlighted source text in the respective documents. As seen in **Figure 3**, standardized names are presented for the compounds; this feature allows the user to systematically assess information that covers different synonyms for a particular drug (one example is Cyclosporine: additional names that will match this drug include ciclosporin, CsA, Neoral, among others). The data can be re-ordered on the fly, so that the results are ordered by frequency (in **Figure 3**: the most common assertions are reported first), alphabetically or on a document-by-document basis.

The dosage column provides dosage information extracted in context. I2E recognizes different expressions for dosage, e.g. "10mg/kg/day", "5mg bid", "5mg rofecoxib per day", "100mg every day", and others.

The information has been extracted from free text in an abstract and converted into structured data in a table. This structure allows the results to be exported into other software packages or processes. For example, from within I2E it is possible to redisplay results directly in Microsoft Excel® and Cytoscape, a network graph viewer.

Pharmacologic Substance	Safety issues	Tissue	Dosage	Doc	Hit
Cyclosporine	Safety issues	Kidney	15 mg/kg/day	17497475	11. of hyperbaric oxygen on cyclosporine-induced nephrotoxicity and oxidative stress in rats. : a control group. Cyclosporine group (15 mg/kg/day) intraperitoneally for 14 days. : a
Mycophenolate Mofetil	Safety issues	Kidney	2 g/day	15041270	3. : day of everolimus or 2 g/day of MMF, plus full-dose cyclosporine (. Concerns over nephrotoxicity led to a protocol amendment.
Amphotericin B	Safety issues	Kidney	0.4 mg/kg/day AMB	15761070	8. Low nephrotoxicity of an effective amphotericin B. for 10 consecutive days with 0.4 mg/kg/day AMB in the form of traditional.
Everolimus	Safety issues	Kidney	everolimus 1.5 mg/day	15041270	3. Concerns over nephrotoxicity led to a protocol amendment. loss to follow-up) were everolimus 1.5 mg/kg/day . 33.7% (65
Lamivudine	Safety issues	Liver	100 mg/day	17293489	1. Lamivudine (100 mg/day) was continued throughout the. Liver Transplantation adverse effects
Alcohol	Safety issues	Liver	20 g/day	15553597	7. Insulin sensitivity and hepatic steatosis in obese subjects with. analyzed 86 obese patients whose alcohol intake was less than 20 g/day and who showed no signs.
Gentamicin	Safety issues	Kidney	100 mg/kg/day	14748758	6. 4-hour tempo) on gentamicin-induced nephrotoxicity in rats. The rats were given gentamicin (100 mg/kg/day) i.p. once a
Sirolimus	Safety issues	Kidney	1 mg/day	15354851	2. primary immunosuppressant in calcineurin inhibitor-induced nephrotoxicity. Sirolimus was started at 1 mg/kg/day with titration over 2 weeks.
Indinavir	Safety issues	Liver	sirolimus 2 mg/day	15899725	2. a 5-mg loading dose of sirolimus participants received sirolimus 2 mg/day for at least 7 days. dropped out because of trimethoprim-sulfamethoxazole-related hepatotoxicity .
Indinavir	Safety issues	Liver	IDV 800/100 mg bid	17263634	4. started zidovudine plus lamivudine plus IDV 800/100 mg bid . was a surprising lack of hepatological side effects during the 6 months of.
Atorvastatin	Safety issues	Liver	10 mg/day	17473378	3. a potential role in statin-related adverse events, and withdrawal of. in patients developing myotoxicity or liver toxicity. Twenty-six patients with hypercholesterolemia received atorvastatin 10 mg/day for 3 months.
Atorvastatin	Safety issues	Liver	atorvastatin 10 mg/day	16731999	1. benefit beyond that afforded by atorvastatin 10 mg/day in patients with stable coronary. in the rates of treatment-related adverse events and persistent elevations in liver enzymes.
Atorvastatin	Safety issues	Liver	atorvastatin 80 mg/day	16731999	1. that intensive lipid-lowering therapy with atorvastatin 80 mg/day provides significant clinical benefit beyond. in the rates of treatment-related adverse events and persistent elevations in liver enzymes.
Acetaminophen	Safety issues	Liver	1000 mg/kg/day	15027815	7. inhibitors in the prevention of hepatotoxicity after paracetamol overdose in rats. The appropriate doses of paracetamol (1000 mg/kg/day) and the inhibitor.

Figure 3. Example I2E smart query results. Highlighting the hits and their context, with links to each document. The user can control the number of results to be returned, along with their preferred results format. Search time, number of documents searched and number of results are also reported (not shown here).

Summary

We have shown here a method of systematic text mining which can be applied to diverse document resources, whether publicly available or proprietary. For this example, we described a strategy for using I2E as a tool in predicting the toxicity profile of potential novel therapeutics.

A subject specific taxonomy has been used to label pharmacologic substances with a controlled vocabulary of primary, preferred terms. Toxicity data has been found and labelled according to the query items for "Liver" and "Kidney" related safety issues.

Dosage information has been extracted directly into a single column in the results table, despite being expressed in a variety of ways in the source texts. Other numerical data can be similarly recognized by I2E, such as concentrations, temperatures or even amino acid and nucleic acid sequence numbering. This latter data can be useful for linking adverse reactions to particular genotypes and protein mutations.

Thus the Linguamatics I2E system can extract detailed information with precision, link information from different documents and uncover hidden knowledge. The rapid speed of search enables fast access directly to new insights. A methodical query strategy, as discussed here, delivers reproducible results and gives confidence that decisions are being supported by high quality information.

Queries can easily be shared and modified enabling a range of end-users to benefit from experts' searches and the power of text mining. Smart queries can be created by information professionals with search expertise. Once built, such queries enable other I2E users to rapidly execute sophisticated searches, which they can tailor to their specific requirements and run on a variety of literature and other text sources.

Acknowledgements

Paul Milligan for query development and application note production.

References

- Milward D, Bjärelund M, Hayes W, Maxwell M, Öberg L, Tilford N, Thomas J, Hale R, Knight S and Barnes J. (2005) Ontology-based interactive information extraction from scientific abstracts. Comparative and Functional Genomics, **6(1-2)**:67-71.
- Milward D, Blaschke C, Neefs J-M, Ott M-C, Verbeeck R and Stubbs A. (2006) Flexible text mining strategies for drug discovery. Proc. Second International Symposium on Semantic Mining in BioMedicine (SMBM), pp. 101-104.
- Milward D and Milligan P. (2007) Text data mining using interactive information extraction. BioLINK SIG Text Mining Workshop, ISMB/ECCB.