Data and Literature Gathering in Chemical Cancer Risk Assessment

Ilona Silins,*† Anna Korhonen,‡ Johan Högberg,† and Ulla Stenius†
†Institute of Environmental Medicine, Karolinska Institutet, Nobels väg 13, S-17177 Stockholm, Sweden
‡Computer Laboratory, University of Cambridge, Cambridge, United Kingdom

(Submitted 2 September 2011; Returned for Revision 8 November 2011; Accepted 9 December 2011)

ABSTRACT

In recent years, chemical cancer risk assessment has faced major challenges: the demand for cancer risk assessment has grown considerably with strict legislation regarding chemical safety, whereas cancer hazard identification has turned increasingly complex due to the rapid development and high publication rate in biomedical sciences. Thus, much of the scientific evidence required for hazard identification is hidden in large collections of biomedical literature. Extensive guidelines have been produced to support cancer risk assessment under these circumstances. We evaluated whether these guidelines support the first, critical step of this task—data and literature gathering—and found that the guidance is vague. We propose ways to improve data and literature gathering for cancer risk assessment and suggest developing a computational literature search and analysis tool dedicated to the task. We describe the first prototype tool we have developed and discuss how it could help to improve the quality, consistency, and effectiveness of cancer risk assessment when developed further. Fully reliable automatic data and literature gathering may not be realistic; the retrieved articles will always need to be examined further by risk assessors. However, our proposal offers a starting point for improved data and literature gathering that can benefit the whole cancer risk assessment process. Integr Environ Assess Manag 2012;8:412–417. © 2010 SETAC

Keywords: Cancer risk assessment Risk assessment Chemical carcinogenesis Text mining Mode of action

INTRODUCTION

Cancer risk assessment of chemicals is a time-consuming and demanding exercise that requires combining scientific knowledge with elaborate literature review. Over the recent years, although the demand for chemical cancer risk assessment has grown, the exercise itself has turned increasingly complex due to the rapid development of molecular biology techniques, the increased knowledge of mechanisms involved in cancer development, and the exponentially growing volume of biomedical literature (e.g., the MEDLINE database of biomedical research articles, affiliated with the US National Library of Medicine, expanded with over half a million references in 2010 [NLM 2011]). These developments have challenged the European Registration, Evaluation, Authorisation and Restriction of Chemical Substances (REACH) implementation in many ways (Hartung 2009). To cope with the situation, an effort has been made to increase the understanding and agreement on basic cancer risk assessment principles via international harmonization projects (e.g., the International Programme on Chemical Safety [IPCS]) and elaborate guidance documents (ECHA 2008a; IARC 2006; USEPA 2005).

DATA AND LITERATURE GATHERING IN CANCER HAZARD IDENTIFICATION—THE STATE OF THE ART

The importance of data and literature gathering for the accuracy, consistency, and efficiency of all risk assessment should not be underestimated. As highlighted by the European Chemicals Agency (ECHA), ''failure to collect all of the available information on a substance may lead to duplicate work, wasted time, increased costs and potentially unnecessary animal use'' (ECHA 2008b). We investigated the support currently available for the first critical step of cancer hazard identification (i.e., the assessment of whether a chemical is capable of causing cancer): data and literature gathering. This step involves 1) identifying all the scientific (e.g., toxicological and epidemiological) data relevant for examining the carcinogenic properties of chemicals, and 2) gathering this data from existing literature. This exercise might be particularly time-consuming when working with new chemicals.

Literature identification and gathering—a time-consuming undertaking

We interviewed 11 experienced risk assessors, most of them with more than 20 years of experience in risk assessment working for various authorities in Sweden, including the Institute of Environmental Medicine at Karolinska Institutet, the Swedish Chemicals Agency, the Scientific Committee on Occupational Exposure Limits (EU), and the Swedish Criteria Group. When asked to identify the most time-consuming step of cancer risk assessment, the majority of them identified data and literature gathering. They explained that this step is time-consuming because data for a single chemical may be scattered across thousands of scientific articles. They conduct data and literature gathering on a largely manual basis, relying only on very basic technical support (e.g., search engines of literature portals) to find the information specified by cancer risk assessment guidelines (ECHA 2008a; IARC 2006; USEPA 2005).

* To whom correspondence may be addressed: Ilona.Silins@ki.se

Published online 19 December 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/ieam.1278

Publicly available cancer risk assessment guidelines

We then investigated the breadth and depth of support offered by these guidelines. We focused our investigation on 2 risk assessment agencies that regulate much of the chemical assessment in the United States and Europe and provide publicly available guidance documents via the internet: the United States Environmental Protection Agency (USEPA) and the ECHA.

The USEPA is the federal governmental agency with the main responsibility of chemical cancer risk assessment in the United States. The USEPA guidelines provide a scientific framework for assessing potential cancer risks from exposures to chemicals in the environment and are also intended to inform decision makers and the public about the recommended procedures. The current version, ''Guidelines for Carcinogen Risk Assessment'' emphasizes the critical role of assessing the Mode of Action (MOA) (USEPA 2005). It gives example sequences of key events that may result in cancer formation (e.g., mutagenesis, increased cell proliferation, and receptor activation). Some scientific evidence are very well documented and associated with a high degree of credibility, whereas others are more hypothetical. The USEPA guidelines are meant to be dynamic, flexible documents that evolve to reflect the current state of science and cancer risk assessment practices.

ECHA is the European Chemicals Agency in Helsinki, Finland, which has been set up to carry out and administer REACH—the recently established European Community regulation on new and existing chemicals and their safe use (EC 1907/2006). ECHA provides guidance documents to inform the industry and risk assessment authorities about the requirements and to assist practical risk assessment work, including cancer risk assessment. We examined documents that cover the data and literature search part and include endpoint-specific guidance for mutagenicity and carcinogenicity (ECHA 2008b). We focused on sections of importance for hazard identification and the assessment of the MOA.

Both the USEPA and ECHA specify the main types of data for cancer risk assessment, including a variable amount of information for each data type. At best, they provide lists of scientific tests that can produce the data in question and that should therefore be searched for in literature. For example, ECHA provides an account of mutagenicity tests that have been a well-established source of cancer risk assessment data (ECHA 2008c). The document gives a detailed list of these tests that can support extensive data and literature search. However, not all the scientific data is specified in such detail. For example, for cell proliferation and apoptosis, the same ECHA document just states that these endpoints are useful in assessing the carcinogenic potential of a substance. Yet these effects (among several other effects not mentioned in the document) are equally important for accurate cancer risk assessment as the well-established mutagenic effects. The vagueness by which nongenotoxic effects are described may be due to the more recent recognition of its importance for carcinogenesis: novel data regarding this intensely studied effect is still more likely to be found in literature describing basic research than in literature typically produced for cancer risk assessment. This is perhaps a systematic problem and reflects a lag phase inherently affecting any regulatory guideline. It may also reflect a recently discussed persistency of the obsolete view that only mutagenic agents can cause cancer (Soto and Sonnenschein 2010).

Current support for literature search and gathering

Although some studies, such as pharmaceutical and pesticide studies, are mainly accessible to governmental agencies, most cancer risk assessment data is now publically available via online literature databases. Services such as the PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), Toxicology Data Network (http://toxnet.nlm.nih.gov/), and the OECD Global Portal to Information on Chemical Substances (http://webnet3.oecd.org/echemportal/) offer standard search engines that enable term or keyword-based literature search of the available data. However, USEPA guidelines provide little support for such literature search. The documents we examined do not cover this practical aspect of cancer risk assessment but rather state that the intention is not to give advice on search strategies. USEPA's Integrated Risk Information System (IRIS) publishes literature searches conducted for some chemicals, including the search terms used and the description of how relevant studies were selected (USEPA 2011). Yet no document explaining these searches or advising on optimal search strategies has been published by USEPA to our knowledge.

In contrast, ECHA provides elaborate advice on literature search, particularly in the recently published ''Guidance on Information Gathering'' document (ECHA 2008b). This document lists potential sources of literature including, for example, in-house files, literature databases, quantitative structure-activity relationship (QSAR) models, along with reviews provided by cancer risk assessment agencies that can provide a useful starting point (e.g., OECD SIDS evaluations, the World Health Organization International Programme of Chemical Safety). Also the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), a nonprofit industry group that e.g., publishes risk assessments is mentioned as a potential source of information. However, no support for optimal search strategies is given.

In summary, although the major cancer risk assessment guidelines examined here provide extensive advice on the range, quality, and ranking of scientific data needed for cancer risk assessment, the advice is not comprehensive or detailed enough to facilitate thorough data gathering. The advice is particularly poor on literature search—an area where risk assessors are not experts and would therefore require particular support. Although ECHA has a dedicated document on information gathering, much of the guidance is too vague to be practical.

Research has revealed conflicting assessments on the same chemical (Rudén 2001). Although such conflicts may be due to variations in national legislation, cancer risk assessment teams, data requirements, and other such variables (Rudén 2006; Seeley et al. 2001), previous research has not considered the possibility that they may also be due to inadequate or a poor choice of data resulting from unsystematic data gathering strategies. Furthermore, the lack of guidance risks efficiency and makes it difficult to conduct data and literature gathering in a systematic and transparent manner.

HOW TO IMPROVE THE STATE OF THE ART

Clearly, there is a need to provide people working with risk assessment with more explicit and up to date guidance on data and literature gathering that can better support chemical cancer hazard identification. This guidance should include a more comprehensive specification of data requirements for cancer risk assessment and hazard identification as well as practical advice on literature gathering, for example, optimal search strategies (e.g., approaches to narrowing down query-based search), ranking of the retrieved data and literature, and keeping track of the entire cancer risk assessment process. It should be provided in such a form that it can be updated whenever the need arises, and it should be linked to regularly maintained online resources that provide the more rapidly changing or developing information, including the specification of the scientific data and the literature resources (e.g., PubMed and cancer risk assessment databases).

However, the existing working practices where risk assessors search for relevant data in the growing body of literature via time-consuming and costly manual means are not optimal. A far better strategy would be to develop a computational literature search tool to support chemical cancer risk assessment. Such a tool can be realistically built using current text mining technology.

Text mining technology for biomedicine

Text mining is a growing field of computer science that develops techniques for automatically retrieving, extracting, and discovering novel information in written texts. The goal is to allow humans to identify required information in literature more efficiently, uncover relationships obscured by the sheer volume of available information, and shift the burden of information overload from the human to the computer. In recent years, biomedical text mining has become increasingly popular due to the great need to provide access to the tremendous body of texts available in biomedical sciences. Technology has been developed to assist various practical tasks, for example, the extraction of databases, dictionaries, summaries, and specific information (e.g., interactions between proteins and genes) from biomedical journal articles and integrated in user-friendly tools (Ananiadou and McNaught 2006; Horn et al. 2004; Muller et al. 2004; Shah et al. 2005). The evaluation of such tools has demonstrated their usefulness for real-world tasks (Cohen et al. 2008; Zweigenbaum et al. 2007).

A text mining tool for assisting risk assessors

Developing text mining technology for chemical cancer hazard identification would enable building a computational tool that conducts much of the initial, time-consuming data and literature gathering automatically. The tool could automatically look for scientific articles that contain data of interest for examining a given chemical, rank and classify these articles based on the evidence they contain, and display the resulting structured data for risk assessors.

We have recently developed a basic prototype version of such a tool (Korhonen et al. 2009; Lewin et al. 2008; Sun et al. 2009). Our tool integrates a taxonomy that specifies scientific data of relevance for cancer risk assessment. The taxonomy focuses on carcinogenic MOA, capturing the current understanding of different processes leading to carcinogenesis. It divides 2 commonly used MOA types, genotoxic and nongenotoxic/indirect genotoxic, further into subtypes (Figure 1). The nongenotoxic/indirect genotoxic MOA categories follow the recently proposed classification of Hattis et al. (2009). Each class of the taxonomy is related to words and phrases that typically indicate the occurrence of the type of scientific data in question. Abstracts concerning 20 selected chemical carcinogens covering different MOAs, e.g., diethylnitrosamine and TCDD, were assigned manually to taxonomy classes and used as training data for machine learning approach that classifies novel literature to the taxonomy (Sun et al. 2009). The method uses linguistic features (e.g., all words in the abstracts) and the annotated material as training data for optimal performance.

According to our evaluation, the results are promising; the tool is capable of classifying novel literature under correct taxonomy nodes with over 95% accuracy (Sun et al. 2009). A screen shot illustrating some of the functions of the tool is provided in Figure 2. As shown in the figure, a user can download a collection of abstracts from PubMed that are then assigned to appropriate taxonomy classes by the tool (Figure 2). The user can thereafter select a taxonomy class and view all the abstracts classified in that class. Because the tool is Web-based, it queries PubMed and enables viewing individual abstracts and articles online.

The tool facilitates both quantitative and qualitative overviews of the data available via PubMed. Patterns, trends, and data gaps can be observed easily in the structured data produced by the tool. To demonstrate how the tool could be used to aid hazard identification, we examined 10 970 MEDLINE abstracts for asbestos, which is a well-known carcinogen. We used the tool to download these abstracts from PubMed and assign them to appropriate taxonomy classes. The tool classified 1300 abstracts as relevant for MOA. Figure 3 shows the distribution of the abstracts for asbestos (in percent) over the different MOA classes. For example, 68% (888) of the abstracts are found under the nongenotoxic/indirect genotoxic class (Figure 3A). Many of these contain evidence for oxidative stress and inflammation (Figure 3B). Asbestos has been shown to induce nongenotoxic effects, such as inflammation and oxidative stress (Donaldson et al. 1989; Nymark et al. 2008), and this can be seen clearly in the distribution of abstracts over the MOA taxonomy, illustrating the usefulness of the tool.

By enabling such analysis, the tool can be used to compare toxicological profiles of unknown chemicals with well-known chemicals to predict their likely properties. It can also be useful for cancer researchers as it can support hypothesis generation for cancer research.

This type of a tool could be developed further in many ways. In addition to improving the coverage (by extending and refining the taxonomy and also to cover other health risks of interest), it could be extended to consider impact factors, citation frequencies, and cross references regarding selected articles. This information could help risk assessors identify more prominent, less important, and incremental studies, as well as studies forming clusters.

The retrieved articles will always need to be examined further by risk assessors. However, a tool based on text mining can help by classifying articles based on the type of evidence they offer for cancer hazard identification and by ranking them according to the amount and strength of the evidence they contain. This saves time and enables risk assessors to identify articles that are likely to be the most useful starting point.

Further research and development is required until a computational literature search tool specifically designed for the needs of cancer risk assessment is available in ''off the shelf'' manner. Once ready, the tool could increase the

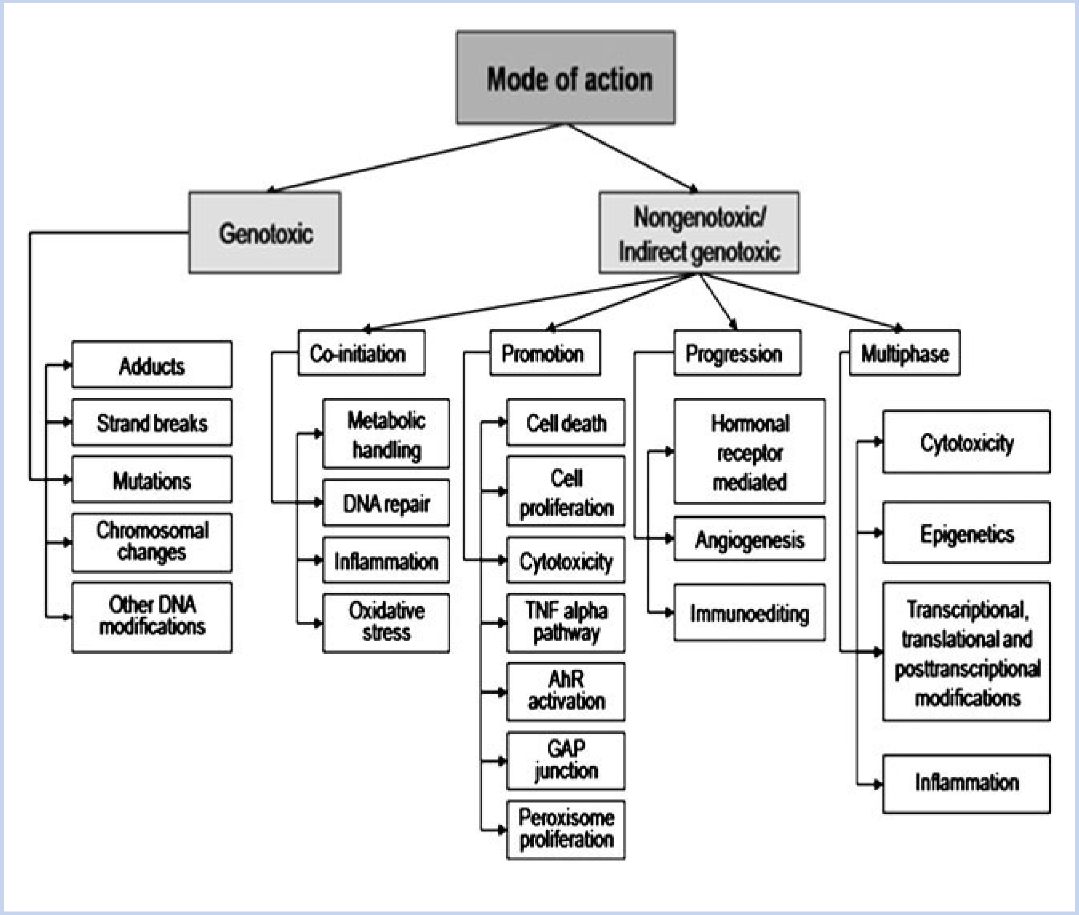
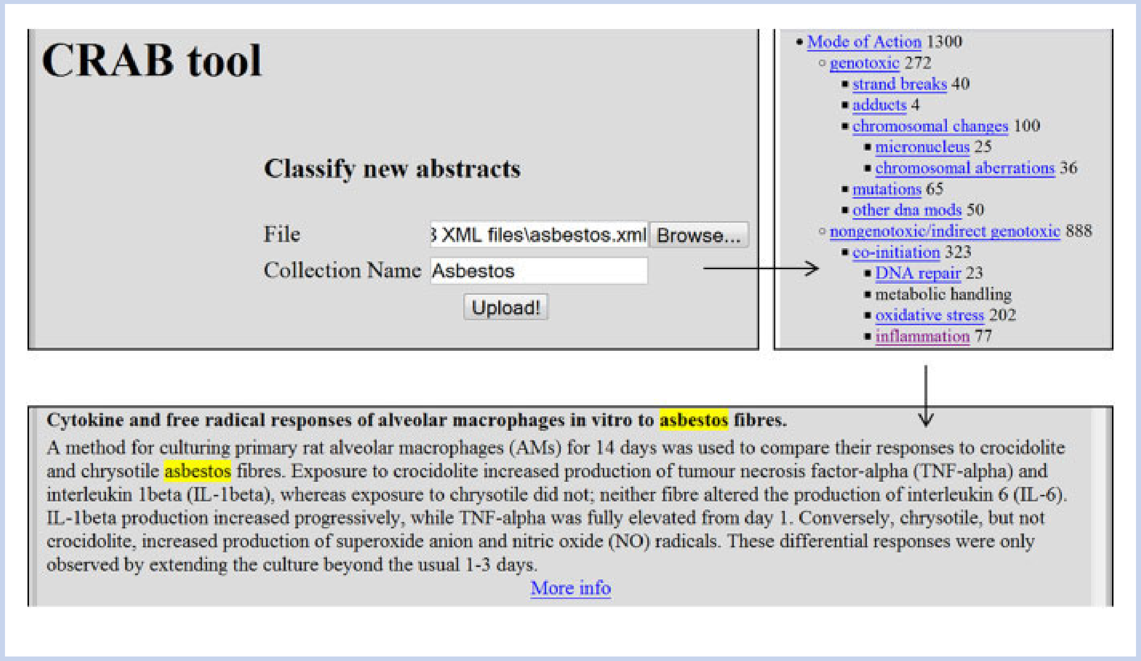


Figure 1. A tentative taxonomy showing genotoxic and nongenotoxic/indirect genotoxic modes of action.



CRAB tool

Classify new abstracts

File 3 XML files\asbestos.xml

Collection Name Asbestos

- Mode of Action 1300
  - genotoxic 272
    - strand breaks 40
    - adducts 4
    - chromosomal changes 100
      - micronucleus 25
      - chromosomal aberrations 36
    - mutations 65
    - other dna mods 50
  - nongenotoxic/indirect genotoxic 888
    - co-initiation 323
      - DNA repair 23
      - metabolic handling
      - oxidative stress 202
      - inflammation 77

Cytokine and free radical responses of alveolar macrophages in vitro to asbestos fibres.

A method for culturing primary rat alveolar macrophages (AMs) for 14 days was used to compare their responses to crocidolite and chrysotile asbestos fibres. Exposure to crocidolite increased production of tumour necrosis factor-alpha (TNF-alpha) and interleukin 1beta (IL-1beta), whereas exposure to chrysotile did not; neither fibre altered the production of interleukin 6 (IL-6). IL-1beta production increased progressively, while TNF-alpha was fully elevated from day 1. Conversely, chrysotile, but not crocidolite, increased production of superoxide anion and nitric oxide (NO) radicals. These differential responses were only observed by extending the culture beyond the usual 1-3 days.

Figure 2. Screen shots of the user interface. The figure shows the index page where the file containing retrieved PubMed abstracts is uploaded, the distribution of asbestos abstracts in MOA classes and a detailed view of 1 abstract found in the ''inflammation'' node.

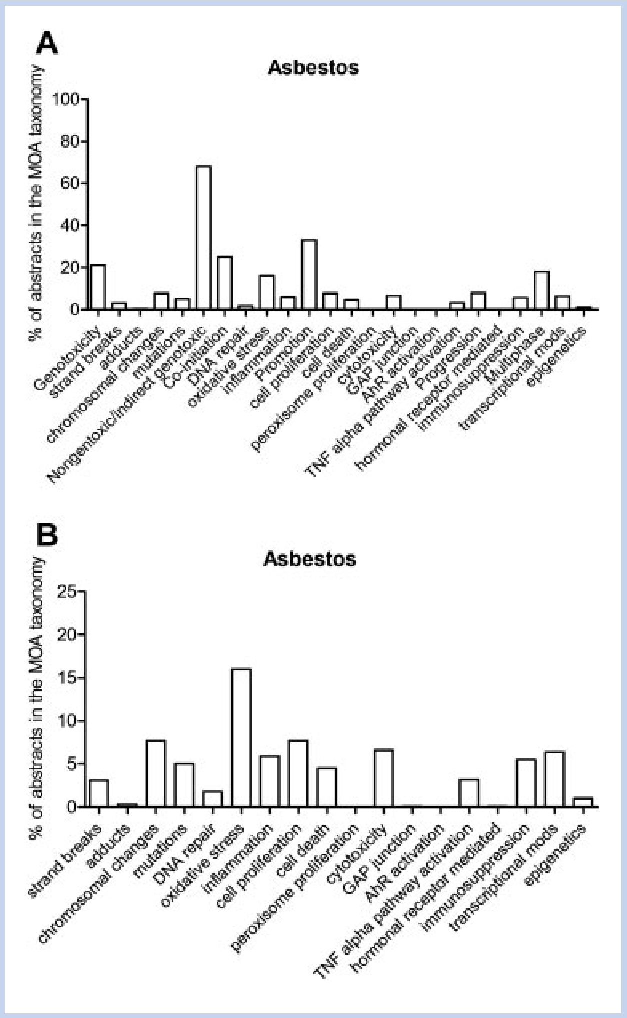


Figure 3. (A) The distribution of 1300 MEDLINE abstracts for asbestos over the MOA taxonomy. (B) A detailed view of some specific MOA nodes.

productivity and efficiency of cancer risk assessment, ensure a more consistent and accurate result, and enable risk assessors to concentrate on what they do best: the expert judgment. It would also help to keep a record of the cancer risk assessment process as recommended by ECHA (ECHA 2008b), providing the practical means to achieve consistency and transparency.

We have developed the prototype tool discussed in this article in the context of our CRAB project (http://www.cl.cam.ac.uk/~alk23/crab/crab.html). The project involves collaboration between the Institute for Environmental Medicine (IEM) at Karolinska Institutet (Sweden) and the University of Cambridge Computer Laboratory (UK). The prototype tool will be made publicly available via the CRAB web page in the near future; meanwhile, it is available on request. In the future, we also plan to develop and extend the tool to cover emerging cancer risk assessment areas and research areas such as early life sensitivity and gender differences in carcinogenesis.

CONCLUSIONS

The circumstances under which cancer risk assessment is conducted have changed dramatically over the past years. Not only the need for chemical assessment (enforced by legislation such as REACH) has increased considerably, pushing existing cancer risk assessment resources to their limits, but also the range of biomedical data of relevance for cancer risk assessment and hazard identification has grown significantly (e.g., by the inclusion of mechanistic data and the importance of defining MOAs). At the same time, the body of potentially relevant literature has grown at a double exponential rate due to the rapidly increasing publication activity in biomedical sciences (e.g., the total volume of published data in PubMed exceeds now 20 million citations from MEDLINE) (http://www.ncbi.nlm.nih.gov/pubmed/). We have presented a proposal that enables improving existing cancer risk assessment work as well developing a computational tool for assisting the cancer risk assessment workflow.

Acknowledgment—Work on this article was funded by the Royal Society (UK), the Medical Research Council (UK), the Swedish Research Council, and the Swedish Council for Working Life and Social Research.

REFERENCES

Ananiadou S, McNaught J. 2006. Text mining for biology and biomedicine. Boston (MA) and London (UK): Artech House Books. 320 p.

Cohen KB, Yu H, Bourne PE, Hirschman L. 2008. Translating biology: Text mining tools that work. In: Pacific Symposium on Biocomputing; 2008 4–8 January; Hawaii, USA. 13: 551–555.

Donaldson K, Brown GM, Brown DM, Bolton RE, Davis JM. 1989. Inflammation generating potential of long and short fibre amosite asbestos samples. *Br J Ind Med* 46:271–276.

[ECHA] European Chemicals Agency. 2008a. Guidance on information requirements and chemical safety assessment. Part B. Hazard assessment. (version 1.1).

[ECHA] European Chemicals Agency. 2008b. Guidance on information requirements and chemical safety assessment. Chapter R.3. Information gathering.

[ECHA] European Chemicals Agency. 2008c. Guidance on information requirements and chemical safety. Chapter R.7a. Endpoint specific guidance. R7.7 mutagenicity and carcinogenicity. p 374–425.

Hartung T. 2009. Toxicology for the twenty-first century. *Nature* 460:208–212.

Hattis D, Chu M, Rahmioglu N, Goble R, Verma P, Kozlak M. 2009. A preliminary operational classification system for nonmutagenic modes of action for carcinogenesis. *Crit Rev Toxicol* 39:97–138.

Horn F, Lau AL, Cohen FE. 2004. Automated extraction of mutation data from the literature: Application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics* 20:557–568.

[IARC] International Agency for Research on Cancer. 2006. Preamble. Monographs on the evaluation of carcinogenic risks to humans. Lyon, France.

Korhonen A, Silins I, Sun L, Stenius U. 2009. The first step in the development of text mining technology for cancer risk assessment: Identifying and organizing scientific evidence in risk assessment literature. *BMC Bioinformatics* 10:303.

Lewin I, Silins I, Korhonen A, Hogberg J, Stenius U. 2008. A new challenge for text mining: Cancer risk assessment. In: Proceedings of the ISMB BioLINK Special Interest Group on Text Data Mining; 2008 20 July. Toronto, Canada: p 1–4.

Muller HM, Kenny EE, Sternberg PW. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2:e309.

[NLM] National Library of Medicine. 2011. [cited 1 September 2011]. Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html

Nymark P, Wikman H, Hienonen-Kempas T, Anttila S. 2008. Molecular and genetic changes in asbestos-related lung cancer. *Cancer Lett* 1–15.

Rudén C. 2006. What influences a health risk assessment? *Toxicol Lett* 167: 201–204.

Rudén C. 2001. The use and evaluation of primary data in 29 trichloroethylene carcinogen risk assessments. *Regul Toxicol Pharmacol* 34:3–16.

Seeley M, Tonner-Navarro LE, Beck BD, Deskin R, Feron VJ, Johanson G, Bolt HM. 2001. Procedures for health risk assessment in Europe. *Regul Toxicol Pharmacol* 34:153–169.

Shah P, Jensen L, Boue S, Bork P. 2005. Extraction of transcript diversity from scientific literature. *PLoS Comput Biol* 1:e10.

Soto AM, Sonnenschein C. 2010. Environmental causes of cancer: endocrine disruptors as carcinogens. *Nat Rev Endocrinol* 6:363–370.

Sun L, Korhonen A, Silins I, Stenius U. 2009. User-driven development of text mining resources for cancer risk assessment. In: Proceedings of the Natural Language Processing in Biomedicine (BioNLP); 2009 4–5 June. Boulder, CO.

[USEPA] US Environmental Protection Agency. 2005. Guidelines for carcinogen risk assessment. Washington (DC): USEPA. EPA/630/P-03/001F.

[USEPA] US Environmental Protection Agency. 2011. Integrated Risk Information System (IRIS). [cited 1 September 2011]. Available from: http://www.epa.gov/ncea/iris/index.html

Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. 2007. New frontiers in biomedical text mining. In: Pacific Symposium on Biocomputing; 2007 3–7 January; Hawaii, USA. 12: 205–208.