

Prioritizing Chemicals for Risk Assessment Using
Chemoinformatics: Examples from the IARC
Monographs on Pesticides

Neela Guha, Kathryn Z. Guyton, Dana Loomis,
and Dinesh Kumar Barupal

<http://dx.doi.org/10.1289/EHP186>

Received: 26 October 2015

Revised: 8 March 2016

Accepted: 28 April 2016

Published: 10 May 2016

Note to readers with disabilities: *EHP* will provide a [508-conformant](#) version of this article upon final publication. If you require a 508-conformant version before then, please contact ehp508@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

Prioritizing Chemicals for Risk Assessment Using Cheminformatics: Examples from the IARC Monographs on Pesticides

Neela Guha, Kathryn Z. Guyton, Dana Loomis, and Dinesh Kumar Barupal

International Agency for Research on Cancer, 150 cours Albert Thomas, 69008 Lyon, France

Address correspondence to Neela Guha and Dinesh Barupal (IARC). E-mail: guhan@iarc.fr
and dinkumar@ucdavis.edu

Short running title: Prioritizing chemicals for risk assessment

Acknowledgments: The IARC Monographs Programme gratefully acknowledges support from the US National Cancer Institute (U01CA033193), the National Institute of Environmental Health Sciences (LAC/IMO/2015/01) and the European Commission (VS/2015/0156).

Author disclaimers: None

Competing financial interest: The authors declare that they have no competing financial interests.

ABSTRACT

Background: Identifying cancer hazards is the first step towards cancer prevention. The IARC Monographs Programme, which has evaluated nearly 1000 agents for carcinogenic potential since 1971, typically selects agents for hazard identification on the basis of public nominations, expert advice, published data on carcinogenicity, and public health importance.

Objectives: Here we present a novel and complementary strategy for identifying agents for hazard evaluation using chemoinformatics, database integration and automated text mining.

Discussion: To inform selection among a broad range of pesticides nominated for evaluation, we identified and screened nearly 6000 relevant chemical structures, thereafter systematically compiled information on 980 pesticides, creating chemical similarity network maps that allowed cluster visualization by chemical similarity, pesticide class, and publicly available information concerning cancer epidemiology, cancer bioassays, and carcinogenic mechanisms. For the IARC Monograph meetings that took place in March and June 2015, this approach supported high priority evaluation of glyphosate, malathion, parathion, tetrachlorvinphos, diazinon, DDT, lindane, and 2,4-D.

Conclusions: This systematic approach, accounting for chemical similarity and overlaying multiple data sources, can be used by risk assessors as well as researchers to systematize, inform and increase efficiency in selecting and prioritizing agents for hazard identification, risk assessment, regulation or further investigation. This approach could be extended to an array of outcomes and agents, including occupational carcinogens, drugs, and foods.

INTRODUCTION

The Monographs Programme of the International Agency for Research on Cancer (IARC) has been instrumental in identifying “environmental” factors that can increase the risk of human cancer. Since its inception in 1971, the Monographs programme has evaluated nearly 1000 agents (as of 2016) and classified them with respect to the strength of scientific evidence that they cause cancer in humans.

The IARC Monographs Programme convenes international Working Groups to identify and classify environmental cancer hazards. The evaluations are based on systematic reviews of epidemiological evidence and cancer bioassay data in experimental animals, with supporting evidence concerning the carcinogenic mechanisms that may act in humans. The sources and extent of human exposure, as well as existing regulations, are also reviewed (IARC preamble 2015). Agents are selected for evaluation by the IARC Monographs Programme through a process that traditionally has relied, to a large extent, on expert recommendations. A public call for nominations of agents is posted on the IARC website and additional nominations are solicited from participating states, the scientific community (including IARC staff), and the general public. An advisory group is then assembled every 5 years to review the nominated agents (Straif et al. 2014) and assign them low, medium or high priority for eventual evaluation. These priority levels reflect the committee’s ranking based on the availability of new data, evidence of carcinogenicity, extent of human exposure, and public health importance (IARC preamble 2015). This method has proved useful for identifying agents of public health importance as priorities for the evaluation of their carcinogenic potential. Following this advice, the Programme selects, groups and orders the priority agents into a series of Monographs, based on a systematic and objective review of the available evidence. Considerations include the priority level assigned by the IARC advisory group, whether and when a compound was last reviewed by IARC and the

potential for the classification to change, usage data (including regulations on use) and the extent of human exposure worldwide, the compound's classification by other agencies and the volume and complexity of informative data that can be reasonably considered during the course of an IARC Monograph meeting. Another important consideration is public health concern, including possible impacts in low- and middle-income countries.

The most recent IARC Advisory Group recommended that the Monographs programme evaluate pesticides. Nominations included particular compounds, chemical classes, related occupations, as well as systematic consideration of cancer-relevant information across a wide range of related exposures. Some pesticides have been identified as potential human carcinogens by authoritative bodies, including IARC, but many are either lacking evaluations or the evaluations may be outdated. Specifically compounds of the organophosphate, organochlorine, triazine, carbamate, dinitroaniline, and pyrethroid classes were accorded priority in 2014 (Straif et al. 2014).

Pesticides include thousands of unique chemical structures distributed across broad chemical and functional classes. Many are chemically or functionally related, but the extent to which they have been studied and the amount of information available from public databases (e.g. PubMed, the Tox21 Program, the PubChem bioactivity assay database) differs markedly across compounds. Given the large amount of data and the structural diversity between compounds, manual review may be prone to incomplete coverage, bias and low efficiency.

Automation of literature mining, integration of electronically available databases and advanced data visualization could be employed as a complimentary approach to systematically incorporate chemical similarity as well as to identify the extent of available information. To address the challenge of appropriately grouping agents and ordering recommended priorities for

hazard assessment, we present a systematic and objective approach using chemoinformatics that has been used to select pesticides for evaluation in recent IARC Monographs (Guyton et al. 2015; Loomis et al. 2015).

METHODS

Overview

A bioinformatics approach was undertaken to systematically assemble and visualize the extent of available information according to chemical similarity across pesticide active compounds. The ranking obtained from the bioinformatics approach was later compiled manually (see Tables 1 and 2) with other important factors considered in selecting priorities (please see the considerations listed in the Introduction above) as presented in Table 1, particularly the assigned priority and the availability of new data to update a previous IARC evaluation.

A first step was to compile a list of all pesticide compounds, which was then organized into chemical similarity network maps. To visualize the availability of data on all pesticides, information by topic area that is considered for an IARC Monograph evaluation (cancer epidemiology, cancer bioassays in animals, mechanistic studies) was then overlaid onto the network maps. Chemical network maps were generated by integrating lists of pesticide compounds with their chemical structure and subsequently mining public databases of the published literature. This process is documented in Figure 1 and detailed in the sections that follow.

Creation of a master list of pesticides

We compiled a master list of pesticide compounds - including chemical name, chemical class and Chemical Abstract Service (CAS) number- from the Kyoto Encyclopedia of Genes and

Genomes (KEGG), MESH Pesticides, EU pesticides, ChEBI pesticides, USEPA Pesticides, and the United States Environmental Protection Agency (USEPA) Toxicity Reference (ToxRef) databases. Chemical structures for each pesticide compound were obtained by linking CAS numbers to PubChem Compound Identifiers (CID) (https://pubchem.ncbi.nlm.nih.gov/search/help_search.html) (see Supplemental Material, Table S1).

Data retrieval from NCBI databases

To assess the scope of the published literature for each pesticide, we searched the titles and abstracts of publications catalogued in PubMed on cancer epidemiology and animal bioassays using the CID, CAS number and various search terms (see Supplemental Material, Table S2). The MeSH term “neoplasm” was used for these searches, since the keyword “cancer” frequently retrieves false positive hits. The papers manually retrieved in our study were also retrieved by the “neoplasms[MeSH]” query, indicating that it covers relevant papers. Searches of ToxRefDB and of PubChem bioassays were also conducted (see Supplemental Material, Table S2). NCBI Eutils and PubChem PUG REST web services were used to systematically query the databases to obtain results of literature and bioassays searches (see Figure 1). Automation for retrieval of data from APIs (Application Programming Interface) was achieved in NodeJS software using JavaScript programming language. To rank pesticides using the chemoinformatics approach, pesticides were sorted by chemical class, the number of publications on cancer and that pesticide overall, the number of cancer epidemiology publications, and the information in ToxRefDB (present or absent).

Chemical similarity network visualization

Network graphs for the chemicals were created using MetaMapp software (Barupal et al. 2012) and visualized in Cytoscape software version 3.1 (U.S. National Institute of General Medical Sciences [http://www.cytoscape.org/release_notes_3_1_1.html]). Individual pesticides are represented as nodes on the chemical similarity maps. Two nodes were linked in the chemical network graph if their Tanimoto similarity score (a coefficient of similarity between two molecules, a measure commonly used in chemoinformatics) was above 0.60, indicating more than 60% chemical similarity. The length of the line connecting the nodes had no meaning itself; it was drawn in reference to the nodes it connected. The node positions within the network maps were controlled by the organic layout algorithm in Cytoscape software which considered a node's degree (the number of connections to a node) and its clustering coefficient (the ratio of the number of actual connections to the total number of possible connections among the node and its neighbors).

A global network of all the pesticides and two focused network graphs of the organophosphorus (OP) and organochlorine (OC) pesticide classes were created. Beyond the KEGG classification, we broadened the pesticide categories for visualization by including those pesticides with at least 1 phosphorous atom or 2 chlorine atoms to the OP and OC pesticide classes, respectively. These network graphs along with the data table used to generate the graph are provided online <http://pesticide.barupal.org/>.

Automated text mining versus directed literature searches

For the top ranking chemicals identified through the chemoinformatics approach, we compared the results from the automated searches to directed PubMed searches. The comparison focused on the cancer epidemiology, as most such studies are found in the published literature. It

also considered any published animal cancer bioassays and studies of key mechanistic evidence (Smith et al. 2015) relating to carcinogenicity of the compound. The manual literature searches and screening were performed using The Health Assessment Workspace Collaborative (hawcproject.org) (Shapiro A 2015).

RESULTS

Creation of a master pesticide list for literature mining

A master list of nearly 6000 pesticide compounds was created from governmental databases, ontologies and databases providing toxicological data on chemicals: KEGG (n = 916 pesticides), MESH Pesticides (n = 451 pesticides), ChEBI (n = 2448 pesticides), USEPA Pesticides (n = 5774 pesticides), EU Pesticides (n = 1318 pesticides) and ToxRefDB (n=474). Entries that were imported from these databases were excluded from the final list if: 1) the structures represented additives such as ethanol 2) they did not have a Chemical Abstracts Service Registry Number (CASRN) or a PubChem Compound Identifier (CID) 3) they were not present in at least 3 of the aforementioned databases and 4) they were compounds that have applications in multiple industries, such as phenol, nicotine, acrolein, and bisphenol A. All the compounds from KEGG, ToxRef, and MESH Pesticide databases were included in the analysis but a number of entries were excluded from the USEPA (n=5024), ChEBI (n=2033), and EU Pesticides databases (n=643). KEGG provided chemical classification information and ToxRefDB provided toxicological data, especially for cancer bioassay data for around 400 selected pesticides. The final list contained 980 pesticide structures (see <http://pesticide.barupal.org/> and <http://pesticide.barupal.org/dataTable.html>).

Selection of pesticides for evaluation in IARC Monograph Volumes 112, 113 and 117

The preceding approach was a starting point for selecting pesticides for evaluation in IARC Monographs volumes 112, 113 and 117 (described below). Using this approach, many of

the top ranked pesticides belonged to the organophosphate (OP) and organochlorine (OC) classes; therefore these pesticide classes were accorded priority. To rank pesticides using the chemoinformatics approach, pesticides were sorted by chemical class, the overall number of publications on cancer and that pesticide, the number of cancer epidemiology publications, and the information in ToxRefDB (present or absent) (see Table 1). The chemical network maps of OPs (Figure 2a) and OCs (Figure 2b) were informative as to the chemical similarity across potential candidates and for identifying related compounds that might be evaluated as a mechanistic class.

In addition to the ranking of OPs and OCs obtained using the chemoinformatics approach (Table 1), several other criteria were considered in order to select a subset of pesticides for evaluation in Monographs 112 and 113 (please see the considerations listed in the Introduction above) as presented in Table 1, particularly the assigned priority and the availability of new data to update a previous IARC evaluation. The volume and complexity of informative data is an important determinant of the number and diversity of agents that can be evaluated in a Monograph meeting. The chemoinformatics approach was therefore useful for visualizing the volume of literature by topic area (cancer epidemiology, animal cancer bioassays, supporting mechanistic evidence). To further refine the list of agents for evaluation, additional directed PubMed searches for epidemiologic and mechanistic data were conducted using standard search strings developed for the IARC Monographs.

Organophosphate pesticides

Based on the preceding criteria, parathion, malathion, diazinon, glyphosate and tetrachlorvinphos emerged as promising candidates for new evaluation in IARC Monograph Volume 112 (Figure 2B). Parathion and malathion, the top-ranked OPs identified from the

chemoinformatics approach, were previously evaluated by the IARC Monographs in 1987 and were then assigned to Group 3 (not classifiable). However, they were later (in 1991 and 2000) classified by the US EPA as potential carcinogens on the basis of positive bioassay data. The availability of newly published epidemiologic studies, particularly for malathion, also supported their selection for re-evaluation in Volume 112. Diazinon, like malathion, was assigned high priority for evaluation by an international advisory group to the IARC Monograph Programme. It ranked fifth by the chemoinformatics approach, had the most cancer epidemiologic studies among the OPs, and had not been previously evaluated by the IARC Monographs Programme.

On the other hand, several candidate agents did not appear to have animal bioassay or other evidence to support a different (e.g., dichlorvos, Group 2B; chlorpyrifos, Group 3) or new (e.g., dimethoate) IARC evaluation. This included several compounds classified by the US EPA in “Group E- Evidence non-carcinogenicity” (terbufos, fonofos, chlorpyrifos and phorate), indicating that animal bioassays had been conducted but none had positive findings [http://npic.orst.edu/chemicals_evaluated.pdf]. An additional factor, noted in Advisory Group recommendations, is that while new epidemiological evidence has emerged it remains incomplete for these agents. Ongoing analyses (e.g., as being conducted by the Agricultural Health Study or AGRICOH), would be important to await before any new or updated evaluation. Thus, these candidates were accorded a lower ranking overall for near-term evaluation.

In contrast, for glyphosate, a recent meta-analysis identified relevant epidemiologic findings (Schinasi and Leon 2014). Additionally, an earlier 1985 classification by the US EPA in Group C indicated the possible availability of pertinent bioassay data (<http://www.epa.gov/iris/subst/0057.htm>). Glyphosate, ranked seventh by the chemoinformatics approach, was assigned medium priority for near-term evaluation by the Advisory Group. The

high production volume of glyphosate leads all OPs and all herbicides, and exposures are widespread, which was another factor in its inclusion among agents in Volume 112. While having some structural similarity to other OPs, glyphosate is toxicologically dissimilar and lacks cholinesterase-inhibiting activity.

Another compelling candidate that emerged from the chemoinformatics approach was tetrachlorvinphos (ranked 13). Tetrachlorvinphos is in current use although overall production volume is low. It was previously evaluated by the IARC Monographs in 1987 and was then assigned to Group 3 (not classifiable). However, tetrachlorvinphos was later (in 2002) classified by the US EPA as a likely human carcinogen based on positive cancer bioassays. Additionally, because tetrachlorvinphos is a direct-acting oxon, in vitro tests for bioactivity might be more informative than for other compounds (e.g. malathion) that require metabolic activation to their oxon forms. These mechanistic considerations together with the positive cancer bioassay findings were the basis for its inclusion in Volume 112. The selection also took into account the overall volume of literature for the other four compounds, with the relatively small size overall of the tetrachlorvinphos literature making it feasible to include.

For the five selected compounds, IARC queried governments and requested public release of government reports on animal cancer bioassay and other relevant data (e.g., genotoxicity) that had been developed by the industry. Direct literature searches identified recently reported epidemiological data, including case–control and cohort studies in the US, Canada, Europe and Sweden. Directed literature searches also identified studies examining relevant carcinogenic mechanisms, including genotoxicity, for both the parent compounds (e.g., malathion, diazinon) and their oxon metabolites. Recent high-throughput data also provided new insights into the extent of biological activity.

In all, these considerations supported the selection of compounds accorded high (malathion, diazinon) or medium (glyphosate) priority for evaluation by the Advisory Group, as well as two others (parathion and tetrachlorvinphos) that were not specifically highlighted in the broad recommendation to evaluate pesticides.

Organochlorine pesticides and 2,4-D

Among the OC pesticides, DDT, lindane, aldrin, and dieldrin were identified as promising candidates for new evaluation according to the criteria described above (Figure 2B). DDT was particularly notable as the pesticide with the largest number of human cancer studies and the largest overall number of PubMed articles retrieved (see table online at <http://pesticide.barupal.org/dataTable.html>). DDT was previously evaluated by IARC in 1991 and had a large number of new studies, lindane had only been evaluated as part of the broader class of hexachlorocyclohexanes and also had new data, while aldrin and dieldrin were last evaluated in 1987 but had relatively few human studies (Table 1). All four are listed as persistent organic pollutants under the Stockholm Convention. DDT and lindane were assigned medium and high priority, respectively, for evaluation by the IARC advisory group, and both had previously been listed as “reasonably anticipated to be a human carcinogen” in the US NTP Report on Carcinogens (USNTP 2014). No additional classifications were identified for aldrin or dieldrin.

A notable literature database was also available for the herbicides 2,4-D and 2,4,5-T (Figure 2B), which were classified in Group 2B by IARC in 1987 as part of the class of chlorophenoxy herbicides (IARC 1987). However, since 2,4,5-T is frequently contaminated with dioxin, which is already classified in IARC Group 1 (IARC 2012), it was not considered further.

As was done for the OP pesticides, directed literature searches identified relevant epidemiology, cancer bioassay and mechanistic studies for the OC pesticides. Additionally, we requested public release of information on cancer bioassays conducted with these compounds that were not available in the public domain.

After considering feasibility, including the unusually large volume of data retrieved for DDT, in addition to the preceding scientific issues, lower ranked aldrin and dieldrin were set aside from the list of potential candidates for later evaluation in IARC Monograph Volume 117 (October 2016), while DDT and lindane were selected for evaluation in Volume 113. 2,4-D was also selected after considering its widespread use and the volume of published literature available (Table 1).

Directed literature searches for pesticides prioritized for evaluation

For the pesticides prioritized for evaluation by IARC in Monograph Volumes 112 and 113 based on the chemoinformatics approach, manual searching and screening of the epidemiological literature was performed using HAWCproject.org. Such manual validation is supported because the size of the published literature—particularly concerning epidemiology or carcinogenic mechanisms—does not always predict the need for a new or updated evaluation. On the other extreme, even a single new well-conducted cancer bioassay could justify further evaluation. Accordingly, the findings of manual screening were compared to the results of the automated searches to determine the relevance of retrieved articles for any resulting evaluation (Table 2; see also Supplemental Material, Table S3).

Generally, the chemoinformatics approach retrieved fewer cancer epidemiology papers than identified through manual searches. For example, for DDT, 190 cancer epidemiology papers were identified through automated searches in PubMed. In comparison, 224 were

identified through targeted manual searches, of which 116 were included after manual review as relevant to an evaluation. Automated searches were not performed concerning the literature on cancer mechanisms as methods to comprehensively identify the broad range of cancer-relevant mechanistic data for potential carcinogens have only recently been advanced (Smith et al. 2015). Nonetheless, targeted searches developed according to the principles outlined by Smith et al (2015) identified a substantial volume of articles on each selected compound. For instance, targeted manual searches and screening of the mechanistic literature for the evaluation of organophosphate pesticides identified relevant publications on malathion (n=370), parathion (n=578), diazinon (n=215), tetrachlorvinphos (n=40) and glyphosate (n=204), respectively. Yet more articles were included for the subsequent evaluation of DDT (n=953), lindane (n=545) and 2,4-D (n=420). Overall, this exercise demonstrated that the chemoinformatics approach provided an efficient and accurate indication of not only the size of the relevant literature, but also identified studies that would be relevant for any resulting evaluation.

DISCUSSION

We illustrate a novel method for the selection of agents for hazard identification that has been applied in the IARC Monographs Programme. Although the data used to construct the chemical network graphs are publicly available, they had not been previously organized in a unified manner that would allow for the simultaneous analysis of the volume of literature on a particular chemical or group of related chemicals. Beyond the KEGG classification, we broadened the pesticide categories for visualization by including those pesticides with at least 1 phosphorous atom or 2 chlorine atoms to the OP and OC pesticide classes, respectively. Doing so enabled us to map pesticides by chemical similarity and include pesticides that may have been missed by pesticide databases or that may have been discarded due to an error in assignment of

chemical class. Accordingly, using a chemoinformatics approach, we were able to integrate information on chemical structure similarity for 980 compounds with the results of systematic, automated text mining of cancer-relevant published information in public databases. The use of web technologies streamlined the integration of information retrieved from different databases/sources and improved efficiency through creating network maps to visualize key chemicals as well as those less studied but chemically related, that may act through a similar mechanism. By enhancing visualization of large-scale public data, our chemoinformatics approach can complement other technologies that employ biomedical text mining strategies to support cancer risk assessment and research (Korhonen et al. 2012).

Using this chemoinformatics approach, pesticides in the organophosphate and organochlorine classes were accorded priority for evaluation in IARC Monograph Volumes 112 and 113: malathion, parathion, diazinon, glyphosate, tetrachlorvinphos (Guyton et al. 2015) and DDT, lindane, and 2,4-D (Loomis et al. 2015). In the resulting evaluations, all of these pesticides were assigned a new or higher IARC classification, reflecting the adequacy of the identified evidence to support these cancer hazard evaluations. In particular, three pesticides previously assigned to Group 3 (not classifiable) were classified in Group 2B (parathion, tetrachlorvinphos) or Group 2A (malathion). Likewise, two others were re-classified from Group 2B to Group 2A (DDT) or Group 1 (lindane). 2,4-D was newly classified in Group 2B; previously IARC had classified the entire class of chlorophenoxy herbicides as Group 2B. Finally, two pesticides, diazinon and glyphosate, both assigned Group 2A, had not been previously classified by IARC. For several of these compounds (including malathion, diazinon, glyphosate, DDT, lindane, and 2,4-D) strong mechanistic evidence supported the resulting

evaluations. In all, these results affirm the utility of the prioritization method for identifying compounds that have evidence warranting new or updated IARC Monograph evaluations.

In addition to the pesticides selected for evaluation in IARC Monograph Volumes 112 and 113, the chemoinformatics approach highlighted several other compounds or compound classes (see <http://pesticide.barupal.org/>). Several of these have been previously evaluated by the IARC Monographs Programme

(http://monographs.iarc.fr/ENG/Classification/latest_classif.php) in Group 2B or higher including trichloroacetic acid (Group 2B, 2014), inorganic arsenic compounds (Group 1, 2012), hexachlorobenzene (Group 2B, 2001) and polychlorophenols (Group 2B, 1999). Others previously assigned to Group 3 and of high use, including atrazine (Group 3, 1999), could be immediately prioritized for re-evaluation together with related compounds (i.e., simazine). Indeed, atrazine was accorded medium priority by the expert advisory group (Straif et al. 2014) based on extensive use and exposures, as well as suspicion of carcinogenicity from newly published information; furthermore, the chemoinformatics method also indicated an extensive literature base. Similarly, diverse compounds currently assigned to IARC Group 3 also emerged (e.g., captan, 1987; methyl bromide, 1999; piperonyl butoxide, 1987), evidently based on information published since the last IARC evaluation. These include some of the most used conventional pesticide active ingredients (e.g., the fumigant methyl bromide, ranked as the 8th (in 2007, 2005 and 2003) or 7th (in 2001) most commonly used conventional pesticide active ingredient by the US EPA (http://www.epa.gov/sites/production/files/2015-10/documents/market_estimates2007.pdf). As noted above for the agents selected for evaluation in Volume 112, chlorpyrifos and other compounds of the organophosphate class may merit re-evaluation following completion and publication of important epidemiological evaluations.

Interestingly, some compounds that emerged as having relevant studies for cancer hazard evaluation have not been previously evaluated or specifically nominated for IARC evaluation (e.g., paraquat).

In general, the chemoinformatics approach retrieved fewer cancer epidemiology papers than identified through the directed literature searches. There are several possible explanations for this discrepancy. Some articles may have been missed because automated searches were of the article title and abstract whereas epidemiology papers sometimes report on multiple pesticides, and specific compounds may not be listed in the title or abstract. The automated searches relied on MeSH annotation using “neoplasms[mesh]” as a more precise search term instead of the keyword “cancer”. This keyword could potentially retrieve irrelevant papers (e.g. that do not describe a laboratory or epidemiological finding on cancer) that MeSH terms would filter. Nonetheless, there may be some delay in assigning publications MeSH annotations and thus more recent but still relevant papers may not be retrieved. The directed searches scanned the full text of publications, enabling identification of publications not retrieved by automated searches using only MeSH terms. A potential limitation of the automated text mining approach is that the search is more specific but less sensitive, sometimes necessitating a manual validation of the literature base to ensure that all relevant publications are captured. However, this limitation affects primarily the later stages of literature retrieval, rather than the initial planning phase. For future efforts using the chemoinformatics approach, we could increase sensitivity of the automated search by broadening the search terms to include key terms as identified from the most informative studies found in a manual search.

Demand for the evaluation of potential chemical hazards is currently increasing, while the resources for testing these chemicals are decreasing (Benigni et al. 2013). The use of long-

term cancer bioassays in animals, which have previously played a fundamental role in the hazard assessment of chemicals, is declining for ethical and practical reasons (e.g. concern for animal welfare, expense, time). Therefore alternative, more cost-effective strategies for predicting the toxicological properties of chemicals, such as (Q)SAR (Quantitative) Structure–Activity Relationships), are being proposed and supported by regulatory initiatives such as REACH (http://ec.europa.eu/environment/chemicals/reach/reach_en.htm) (Benigni et al. 2013;van et al. 2009). These approaches capitalize on the wealth of data already captured in publicly available databases.

By employing technological advances in bioinformatics and computational toxicology, we demonstrate that the use of chemoinformatics is a powerful and complementary approach for prioritizing chemicals for risk assessment. This approach could be further extended to support prediction of emerging risks and informed substitution of hazardous chemicals by “safer” alternatives (Jacobs et al. 2016) wherein bioinformatics approaches could compare compounds that are not yet tested, but structurally similar, to agents already classified for their carcinogenic potential by the IARC Monographs Programme. Epidemiologists and other researchers assessing associations between numerous chemicals and outcomes may also be able to employ this strategy to identify agents for further investigation when designing large-scale studies of human health. Since national health agencies use the information from the IARC Monographs as scientific support for their actions to prevent or reduce exposure to potential carcinogens, efficiently prioritizing agents for risk assessment and predicting emerging hazards are important steps towards protecting public health.

CONCLUSION

Using a novel chemoinformatics approach, we integrated information on chemical structure similarity with the results of systematic, automated text mining of cancer-relevant information in public databases to select chemical agents for hazard identification. We demonstrate this as an efficient method for grouping of chemicals within class in selecting agents for hazard evaluation in the IARC Monographs. This systematic approach, accounting for chemical similarity and overlaying multiple data sources, can be used to systematize, inform and increase efficiency in selecting and prioritizing agents for hazard identification, risk assessment and regulation or further investigation. Further, by overlaying new chemicals on to a network map of agents already classified by the IARC Monographs, emerging risks and potential cancer hazards (e.g occupational carcinogens, drugs, environmental pollutants, nutritional compounds) might be identified. This innovation could be extended to an array of outcomes and agents and may prove particularly useful to national regulatory agencies for prioritizing agents for risk assessment and regulation.

REFERENCES

- Barupal DK, Haldiya PK, Wohlgemuth G, Kind T, Kothari SL, Pinkerton KE, et al. 2012. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics* 13: 99.
- Benigni R, Bossa C, Battistelli CL, Tcheremenskaia O. 2013. IARC classes 1 and 2 carcinogens are successfully identified by an alternative strategy that detects DNA-reactivity and cell transformation ability of chemicals. *Mutat Res* 758: 56-61.
- Guyton KZ, Loomis D, Grosse Y, El GF, Benbrahim-Tallaa L, Guha N, et al. 2015. Carcinogenicity of tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate. *Lancet Oncol* 16: 490-491.
- IARC. 1987. Overall evaluations of carcinogenicity: an updating of IARC Monographs volumes 1 to 42. *IARC Monogr Eval Carcinog Risks Hum Suppl* 7: 1-440.
- IARC. 2012. Chemical agents and related occupations. *IARC Monogr Eval Carcinog Risks Hum* 100: 9-562.
- IARC preamble. 2015.
- Jacobs MM, Malloy TF, Tickner JA, Edwards S. 2016. Alternatives Assessment Frameworks: Research Needs for the Informed Substitution of Hazardous Chemicals. *Environ Health Perspect* 124: 265-280.
- Korhonen A, Seaghdha DO, Silins I, Sun L, Hogberg J, Stenius U. 2012. Text mining for literature review and knowledge discovery in cancer risk assessment and research. *PLoS One* 7: e33427.
- Loomis D, Guyton K, Grosse Y, El GF, Bouvard V, Benbrahim-Tallaa L, et al. 2015. Carcinogenicity of lindane, DDT, and 2,4-dichlorophenoxyacetic acid. *Lancet Oncol* 16: 891-892.
- Schinasi L, Leon ME. 2014. Non-Hodgkin lymphoma and occupational exposure to agricultural pesticide chemical groups and active ingredients: a systematic review and meta-analysis. *Int J Environ Res Public Health* 11: 4449-4527.
- Shapiro A. 2015. Health Assessment Workspace Collaborative (HAWC). 2013. [accessed 1 August 2015].
- Smith MT, Guyton KZ, Gibbons CF, Fritz JM, Portier CJ, Rusyn I, et al. 2015. Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environ Health Perspect*.

Straif K, Loomis D, Guyton K, Grosse Y, Lauby-Secretan B, El Ghissassi F, et al. 2014. Future priorities for the IARC Monographs. *The Lancet Oncology* 15: 683-684.

USNTP. 2014. Report on Carcinogens, Thirteenth Edition. Available:
<http://ntp.niehs.nih.gov/pubhealth/roc/roc13/> .

van LK, Schultz TW, Henry T, Diderich B, Veith GD. 2009. Using chemical categories to fill data gaps in hazard assessment. *SAR QSAR Environ Res* 20: 207-220.

Table 1. High ranking organophosphate, organochlorine and chlorophenoxy pesticides identified using a chemoinformatics approach.

Name	Rank	PubMed cancer hits	PubMed human cancer hits	IARC Advisory Group Priority	Other classifications	Usage notes	Prior IARC classification (year)	2015 IARC classification
Organophosphates								
Parathion	1	42	6	-	US EPA Group C (1991) ^a	Restricted ^b	3 (1987)	2B
Malathion	2	40	12	High	US EPA Suggestive (2000) ^a	High ^c	3 (1987)	2A
Chlorpyrifos	3	38	14	Medium	US EPA Group E (1993) ^a	High ^{d,e}	3 (1987)	
Dichlorvos	4	35	12	-	US EPA Suggestive (2000) ^a	Some current uses ^f	2B (1991)	
Diazinon	5	30	16	High	US EPA Not likely (1997) ^a	High ^c	2A (2015)	2A
Glyphosate	7	21	9	Medium	US EPA Group C (1985), Group E (1991) ^a	High ^{d,e}	None	2A
Tetrachlorvinphos	13	6	1	-	US EPA Likely (2002) ^a	Currently used	3 (1987)	2B

^a For a description of US EPA cancer classifications, see <http://www.epa.gov/pesticides/health/cancerfs.htm#terms> .

^b Severely banned or restricted for health or environmental reasons (Rotterdam Convention, Annex III)

^c Among the most commonly used OP insecticides in all US market sectors (2001 to 2007), malathion and chlorpyrifos are listed 1 or 2; diazinon is listed 3 (2001, 2003) or 8 (2005, 2007) (<http://www.epa.gov/pesticides/pestsales/>) .

^d Chlorpyrifos and glyphosate are among the most commonly used conventional pesticide active ingredients in the US (<http://www.epa.gov/pesticides/pestsales/>).

^e Glyphosate was the most commonly used conventional pesticide in the agricultural market sector from 2001 to 2007; in this same period chlorpyrifos ranked 13 (2003), 14 (2007), or 15 (2001, 2005) (<http://www.epa.gov/pesticides/pestsales/>).

^f Many domestic and other uses of dichlorvos in the US have been discontinued (<http://www.gpo.gov/fdsys/pkg/FR-1995-04-19/pdf/95-9166.pdf>)

Organochlorines								
DDT	1	494	190	Medium	POP ^g , RoC-RA ^h	Restricted ^b	2B (1991)	2A
Lindane	2	189	51	High	POP ^g , RoC-RA ^h ; US EPA Suggestive (2001) ^a	Restricted ^b	2B ⁱ (1987)	1
Dieldrin	3	151	57	-	POP ^g	Restricted ^b	3 (1987)	
Aldrin	7	56	25	-	POP ^g	Restricted ^b	3 (1987)	
Chlorophenoxy								
2,4-Dichlorophenoxy acetic acid	1	145	84	-	US EPA Group D (2004) ^a	Currently used	None	2B

^g POP, Listed as a persistent organic pollutant under the Stockholm Convention

^h RoC-RA, Listed as “reasonably expected to be a human carcinogen” in the US Report on Carcinogens

Table 2. Cancer epidemiology literature retrieved through automated mining (chemoinformatics) and manual PubMed searches for pesticides evaluated in IARC Monographs 112 and 113.

Pesticide	Chemoinformatics	Manual searches		
	Retrieved	Retrieved	Included	Excluded
Malathion	12	80	28	52
Parathion	6	12	9	3
Diazinon	16	39	22	17
Tetrachlorvinphos	1	4	4	0
Glyphosate	9	50	19	31
DDT	190	224	116	107
Lindane	51	46	22	24
2,4-D	84	76	62	11

Figure Legends

Figure 1. Overall scheme of the chemoinformatics approach for data retrieval and visualization for the prioritization of pesticides for the evaluation of their carcinogenic potential. A. PubMed Cancer All; B. PubMed Cancer Epidemiology; C. PubMed animal cancer bioassays; D. ToxRefDB carcinogenicity; E. Chemical Similarity Scores. See Supplemental Table S2 for a description of the search terms.

Figure 2. Focused visualization of the chemical similarity network maps for A) organophosphorus and other pesticides with at least 1 phosphorous atom and B) organochlorines and other pesticides with more than 2 chlorine atoms.

Individual pesticides are represented as nodes on the chemical similarity maps. The node size is proportional to the number of publications overall on a pesticide and cancer: larger nodes represent more publications. The node border width represents the number of publications on epidemiology, cancer and the pesticide: a thicker border represents more papers. The node color also represents the number of publications on epidemiology, cancer and the pesticide: red represents the highest count of publications. The node shape indicates whether results for a particular pesticide were available in the ToxRefDB database (circle = absent; square = present). The node border color represents the KEGG pesticide classification. [Greater detail on the colors used as well as the associated table describing the information in the figures can be found online http://pesticide.barupal.org/.](http://pesticide.barupal.org/)

Two nodes are linked by a line if their Tanimoto similarity score is above 0.60 (hence they are more than 60% chemically similar). The length of the line connecting the nodes has no meaning itself; it is drawn in reference to the nodes it is connecting. The node positions within the network maps are controlled by the organic layout algorithm in Cytoscape software which considers a node's degree (the number of connections to a node) and its clustering coefficient (the ratio of the number of actual connections to the total number of possible connections among the node and its neighbors).

The session file that can be opened in Cytoscape for zoom-in and focused visualization is located online <http://pesticide.barupal.org/>.

Figure 1.

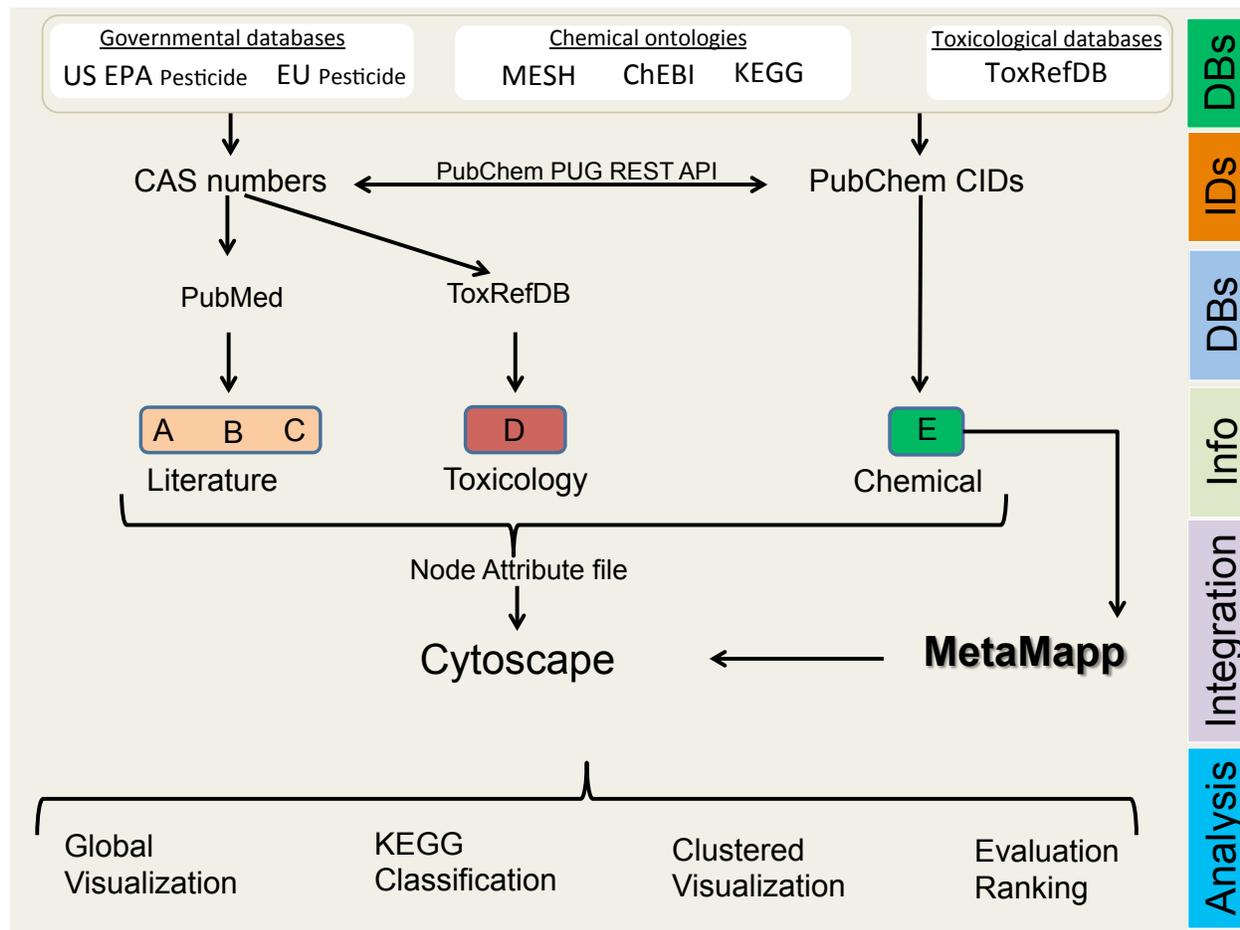


Figure 2.

