A Section 508–conformant HTML version of this article
is available at https://doi.org/10.1289/EHP6702.

Application of Text Mining in Risk Assessment of Chemical Mixtures: A Case Study of Polycyclic Aromatic Hydrocarbons (PAHs)

Imran Ali,[1] Kristian Dreij,[1] Simon Baker,[2] Johan Högberg,[1] Anna Korhonen,[2] and Ulla Stenius[1]

[1]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[2]Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, UK

Background: Cancer risk assessment of complex exposures, such as exposure to mixtures of polycyclic aromatic hydrocarbons (PAHs), is challenging due to the diverse biological activities of these compounds. With the help of text mining (TM), we have developed TM tools—the latest iteration of the Cancer Risk Assessment using Biomedical literature tool (CRAB3) and a Cancer Hallmarks Analytics Tool (CHAT)—that could be useful for automatic literature analyses in cancer risk assessment and research. Although CRAB3 analyses are based on carcinogenic modes of action (MOAs) and cover almost all the key characteristics of carcinogens, CHAT evaluates literature according to the hallmarks of cancer referring to the alterations in cellular behavior that characterize the cancer cell.

Objectives: The objective was to evaluate the usefulness of these tools to support cancer risk assessment by performing a case study of 22 European Union and U.S. Environmental Protection Agency priority PAHs and diesel exhaust and a case study of PAH interactions with silica.

Methods: We analyzed PubMed literature, comprising 57,498 references concerning priority PAHs and complex PAH mixtures, using CRAB3 and CHAT.

Results: CRAB3 analyses correctly identified similarities and differences in genotoxic and nongenotoxic MOAs of the 22 priority PAHs and grouped them according to their known carcinogenic potential. CHAT had the same capacity and complemented the CRAB output when comparing, for example, benzo[*a*]pyrene and dibenzo[*a,l*]pyrene. Both CRAB3 and CHAT analyses highlighted potentially interacting mechanisms within and across complex PAH mixtures and mechanisms of possible importance for interactions with silica.

Conclusion: These data suggest that our TM approach can be useful in the hazard identification of PAHs and mixtures including PAHs. The tools can assist in grouping chemicals and identifying similarities and differences in carcinogenic MOAs and their interactions. https://doi.org/10.1289/EHP6702

Introduction

Hazard identification is the first step in cancer risk assessment and requires a thorough review of the literature for generating information on the carcinogenic potential of chemicals. An efficient and reliable literature review is crucial to properly assess impact on human health (Goodman and Lynch 2017). However, owing to the exponential growth and complexity of the scientific literature, manual evaluation has become extremely challenging and time consuming (Korhonen et al. 2012). While considering human exposures to multitude chemical carcinogens and the pace of current risk assessment practices, novel approaches need to be introduced to evaluate cancer risks associated with these chemicals. Therefore, we have developed text-mining (TM) tools for literature review and to identify mechanistic information for cancer risk assessment of chemicals (Korhonen et al. 2012; Silins et al. 2012). Although prior research has produced some tools to support practical tasks such as literature curation and development of semantic databases in biomedicine, none of these are focused on chemical cancer risk assessment (Harmston et al. 2010; Rebholz-Schuhmann et al. 2012; Zhu et al. 2013).

Based on TM technology, we developed an open access Cancer Risk Assessment using Biomedical literature (CRAB) tool that automatically organizes and classifies literature according to carcinogenic modes of action (MOAs) (Ali et al. 2016; Korhonen et al. 2009, 2012; Silins et al. 2012). Our evaluations have shown that the automatic classification results generated by these tools are very accurate, often between 95% and 99% (Korhonen et al. 2012). In addition, the CRAB tool can be used to reveal similarities and differences between chemicals by, for example, identifying similar MOAs shared by chemicals in mixtures. We have employed the tool in practice, including studies of MOA analysis for groups of chemicals. For example, in a study of polychlorinated biphenyls (PCBs), we analyzed literature profiles of dioxin-like- and non-dioxin–like PCBs and compared them with those of their indicator PCBs (Ali et al. 2016). A comparison of carcinogenic MOAs identified by the CRAB tool with a manual evaluation of PCBs found similar evidence on key characteristics (Smith et al. 2016), associated with carcinogenicity of these compounds. We also investigated carcinogenic MOAs of pesticides in fruits and identified that CRAB-generated MOA profiles of many pesticides were similar, containing evidence for both genotoxic and nongenotoxic MOAs (Silins et al. 2014). Together, these results suggest that the CRAB tool can be of great use in effectively identifying groups of chemicals with similar MOAs and to evaluate the relevancy of a reference compound for a group of chemicals.

In addition to the CRAB tool, we have developed a Cancer Hallmarks Analytics Tool (CHAT) (Baker et al. 2017), primarily intended to assist cancer researchers. CHAT provides detailed analyses of literature by automatically organizing and classifying literature on the basis of the evidence on cancer hallmarks and associated processes. CHAT consists of 37 distinct categories, which specify the biological processes underlying each of the original eight cancer hallmarks and two enabling characteristics (Hanahan and Weinberg 2000, 2011). CHAT analyses provide mechanistic information on the biological processes involved in cancer development.

Exposures to chemicals occur in mixtures, and mixture risk assessment is challenging (Gwinn et al. 2017). One such group of chemicals comprises the polycyclic aromatic hydrocarbons (PAHs), a family of more than 1,500 compounds with diverse structural features, that is, ring configurations and substitutions (NTP 2012). Based on the evidence on their carcinogenicity, the International Agency for Research on Cancer (IARC) has

Address correspondence to Imran Ali, Box 210, Institute of Environmental Medicine, Karolinska Institutet, SE-17177 Stockholm, Sweden. Email: imran.ali@ki.se

Supplemental Material is available online (https://doi.org/10.1289/EHP6702).

The authors declare they have no actual or potential competing financial interests.

Received 19 December 2019; Revised 7 May 2021; Accepted 10 May 2021; Published 24 June 2021.

classified the PAH benzo[*a*]pyrene (B[*a*]P) as a human carcinogen and several other PAHs as possible or probable human carcinogens (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans 2010). Although the evaluation of carcinogenicity of PAHs has been made based on their mutagenic and tumor-initiating activities, data are limited on their tumor-promoting activity (Asada et al. 2005; Sakai et al. 2010). Recently it was suggested that some of the PAHs with relatively low carcinogenic potential exhibit strong tumor-promoting activity (Misaki et al. 2016). Importantly, many complex mixtures containing PAHs have been classified as human carcinogens—including coal tar, diesel exhaust, and coke oven emission—and are of serious health concern, not only for the general population, but also for those in occupational settings (IARC 2016). Although sufficient data are available to support the carcinogenicity of some PAH mixtures in humans, there is a general lack of epidemiological data for individual PAHs, for example, B[*a*]P (Baan et al. 2009).

Cancer risk assessment of PAHs is principally based on one of two approaches. One is based on B[*a*]P as a surrogate marker for all PAHs and thus solely considers the environmental levels of B[*a*]P to assess risks. This approach is used for assessing airborne PAHs within the European Union (WHO 2000). The other approach considers the contribution from other PAHs by applying potency factors that describe the relative carcinogenic potency to B[*a*]P (RPFs) and is used by, for example, the U.S. Environmental Protection Agency (EPA) for assessing environmental PAH pollution (U.S. EPA 1993). Although the latter approach includes multiple PAHs, only a limited number of PAHs have been assigned a RPF, and an extended list is currently under evaluation (U.S. EPA 2010a). A potentially serious limitation of both approaches is that they do not allow for taking interaction effects into account. A number of studies have shown that PAH mixtures can induce nonadditive effects, resulting in antagonistic or synergistic genotoxic or carcinogenic effects (Jarvis et al. 2014). Therefore, the development of new methodology that will optimize the possibility of taking advantage of published scientific databases is highly motivated.

In this article, we present a case study of 22 U.S. EPA and EU priority PAHs and complex PAH mixtures that are of concern for the general population and in occupational settings. We used the CRAB3 tool (http://crab3.lionproject.net/)—the latest iteration of the CRAB tool—and CHAT (http://chat.lionproject.net/) to analyze PAH literature and to evaluate the usefulness of these tools for hazard identification. CHAT analyses complement those of the CRAB3 tool, and to the best of our knowledge, for the first time they were used together.

Material and Methods

Research Question and Literature Gathering

We used the TM-based tools CRAB3 and CHAT to assess the usefulness of these tools to support cancer risk assessment and how these tools can be used together to *a*) identify the primary carcinogenic MOA of a group of chemicals, *b*) group these chemicals according to their carcinogenic potential, and *c*) identify possible interaction effects within and between complex mixtures. Our motivation for choosing PAHs for a case study was mentioned above. To answer these questions, we analyzed PubMed literature (abstracts) concerning 22 priority PAHs listed by the European Union and U.S. EPA (EC 2002; Jarvis et al. 2014; U.S. EPA 2010b), some complex mixtures containing these PAHs, and crystalline silica. The lists of PAHs and PAH mixtures evaluated are shown in Tables 1 and 2, respectively. Literature on PAHs were further grouped based on their relative potency factor (RPF) values (Table 1), namely, Group 1, with low potency PAHs (RPF 0.001–0.03); Group 2, with moderate potency PAHs (RPF 0.08–0.8); and Group 3, with high potency PAHs (RPF 1–100, excluding B[*a*]P) and compared between groups and with the MOA profile of B[*a*]P (RPF = 1). We performed the literature analyses by using the CRAB3 tool and CHAT separately given that these tools are not coupled together, and the data cannot be automatically transferred between them. The tools and associated analyses are described in more detail below.

Table 1. Range of RPF values for the 22 priority PAHs listed by the U.S. EPA and European Union (Jarvis et al. 2014).

PAHs	RPF (range)
Naphthalene (NAP)	0–0.001
Acenaphthylene (ACY)	0.001
Acenaphthene (ACE)	0.001
Fluorene (FLO)	0.0005–0.001
Phenanthrene (PHE)	0 to <0.01
Pyrene (PYR)	0–0.001
Anthracene (ANT)	0–0.01
Benzo[*ghi*]perylene (B[*ghi*]Per)	0.009–0.03
Fluoranthene (FLA)	0.001–0.08
Benzo[*k*]fluoranthene (B[*k*]F)	0.03–0.1
Chrysene (CHR)	0.01–0.17
Benz[*a*]anthracene (B[*a*]A)	0.005–0.2
Cyclopenta[*cd*]pyrene (CP[*cd*]P)	0.012–0.4
Benzo[*j*]fluoranthene (B[*j*]F]	0.045–0.52
Benzo[*b*]fluoranthene (B[*b*]F)	0.1–0.8
Dibenzo[*a,e*]pyrene (DB[*ae*]P)	0.2–1
Benzo[*a*]pyrene (B[*a*]P)	1
5-Methylchrysene (meCHR)	7
Dibenzo[*a,h*]anthracene (DB[*ah*]A)	0.1–10
Dibenzo[*a,h*]pyrene (DB[*ah*]P)	0.9–11
Benzo[*c*]fluorene (B[*c*]F)	20
Dibenzo[*a,l*]pyrene (DB[*al*]P)[a]	1–100

Note: EPA, Environmental Protection Agency; PAHs, polyaromatic hydrocarbons; RPF, relative potency factor.
[a]Also known as dibenzo(*def,p*)chrysene.

CRAB3 Analyses

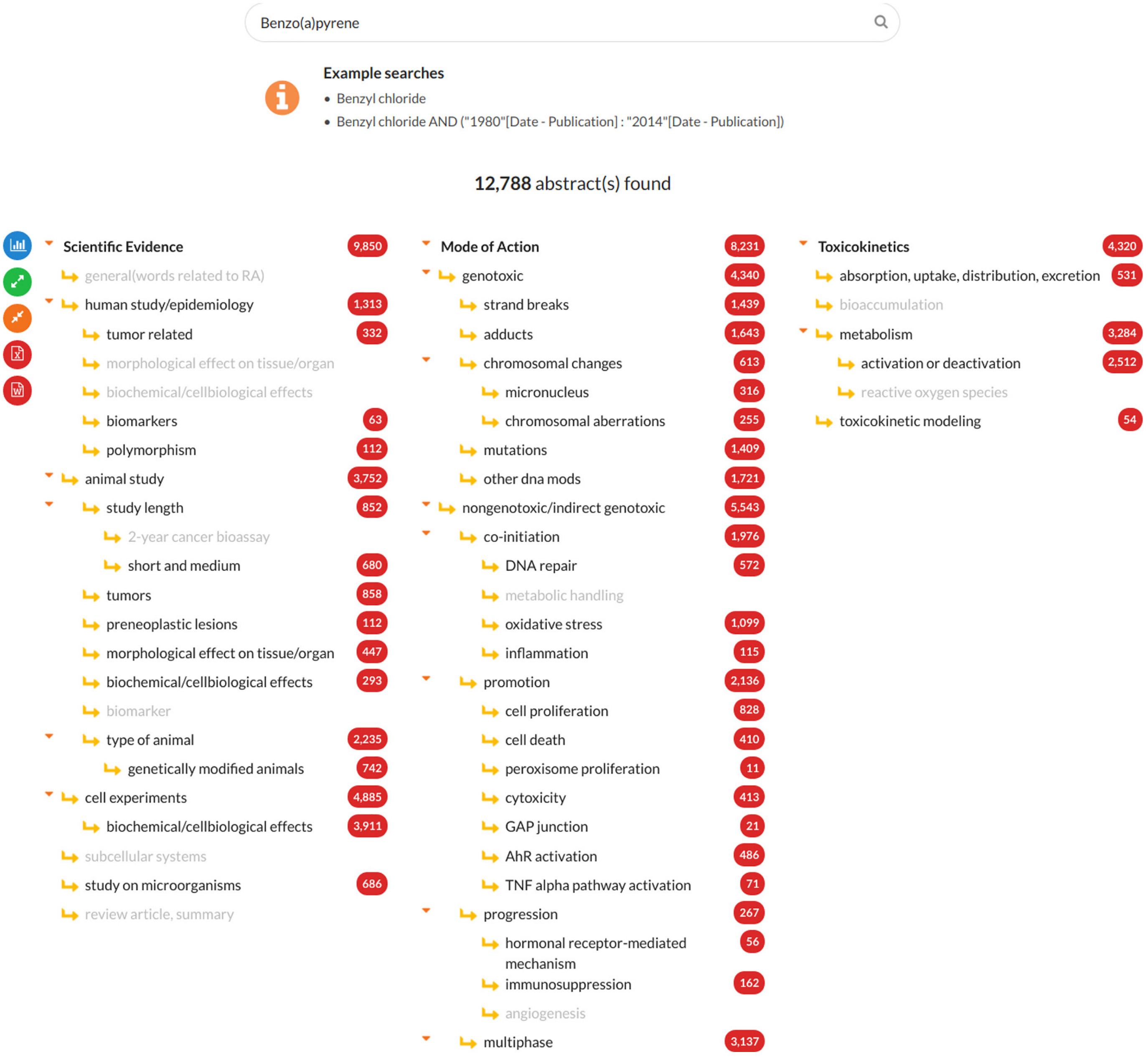
The CRAB3 tool analyzes literature on the basis of the textual content of each title and abstract using natural language processing followed by semantical classification using supervised machine learning. The classification is based on type of scientific evidence (human/epidemiology study, animal study, cell experiments and study on microorganisms) and according to a carcinogenic MOA taxonomy followed by an automated statistical analysis of the classified literature (Korhonen et al. 2012). The MOA taxonomy was developed based on the genotoxic and nongenotoxic MOA classification by Hattis et al. (2009) A unique feature of CRAB3 is that it identifies argumentative zones in the abstract text, a scheme of information structure that describes the rhetorical progression in scientific articles, denoting the scientific discourse, namely, "Objective," "Background," "Method," "Results," "Conclusion," and "Future Work" sections that appear in the abstract (Guo et al. 2014). Furthermore, for comparison analysis between two chemicals, the built-in feature "analyze data" can be used to visualize similarities and differences in the MOA profiles (Figure S1).

For the current MOA analyses of PAHs, we used the search query function of CRAB3 by simply adding the name of individual PAHs (Table 1) to identify PubMed literature (available in PubMed March 2019) on the 22 priority PAHs. An example is shown in Figure 1. The comparison analysis function was used for pairwise comparison of MOA profiles between different unsubstituted and substituted PAHs. In addition, we queried the following complex PAH mixtures: diesel exhaust, coal tar, coke oven emission, and crystalline silica. For pyrene analysis, we processed the CRAB3 data manually and excluded literature on

Table 2. An overview of the literature on 22 priority PAHs, complex PAH mixtures and crystalline silica available in PubMed (March 2019), and the percentages of the literature classified as relevant to scientific evidence and mode of action (MOA) taxonomies by CRAB3 (Korhonen et al. 2012).

CRAB evaluation	Total PubMed references (*n*)	References in scientific evidence [*n* (%)]	References in MOAs [*n* (%)]
22 Priority PAHs	53,944	22,657 (42)	22,031 (41)
Benzo[*a*]pyrene	11,887	9,236 (78)	7,797 (65)
Diesel exhaust	2,181	1,300 (56)	1,207 (55)
Coal tar	892	465 (52)	320 (36)
Coke oven emission	481	375 (78)	246 (51)
Crystalline silica	939	592 (63)	369 (39)

Note: CRAB3 was used to search PubMed for literature on 22 priority PAHs, complex PAH mixtures, and crystalline silica. Shown in the table are the total number of references the CRAB3 tool identified, the number of references with available information on study design (e.g., human studies, animal studies, cell experiments), and the number of references with information on carcinogenic modes of action. CRAB3, Cancer Risk Assessment using Biomedical literature tool 3; PAH, polycyclic aromatic hydrocarbon.



Benzo(a)pyrene

Example searches

- Benzyl chloride
- Benzyl chloride AND ("1980"[Date - Publication] : "2014"[Date - Publication])

12,788 abstract(s) found

- **Scientific Evidence** 9,850
  - general(words related to RA)
  - human study/epidemiology 1,313
    - tumor related 332
    - morphological effect on tissue/organ
    - biochemical/cellbiological effects
    - biomarkers 63
    - polymorphism 112
  - animal study 3,752
    - study length 852
      - 2-year cancer bioassay
      - short and medium 680
    - tumors 858
    - preneoplastic lesions 112
    - morphological effect on tissue/organ 447
    - biochemical/cellbiological effects 293
    - biomarker
    - type of animal 2,235
      - genetically modified animals 742
  - cell experiments 4,885
    - biochemical/cellbiological effects 3,911
  - subcellular systems
  - study on microorganisms 686
  - review article, summary

- **Mode of Action** 8,231
  - genotoxic 4,340
    - strand breaks 1,439
    - adducts 1,643
    - chromosomal changes 613
      - micronucleus 316
      - chromosomal aberrations 255
    - mutations 1,409
    - other dna mods 1,721
  - nongenotoxic/indirect genotoxic 5,543
    - co-initiation 1,976
      - DNA repair 572
      - metabolic handling
      - oxidative stress 1,099
      - inflammation 115
    - promotion 2,136
      - cell proliferation 828
      - cell death 410
      - peroxisome proliferation 11
      - cytoxicity 413
      - GAP junction 21
      - AhR activation 486
      - TNF alpha pathway activation 71
    - progression 267
      - hormonal receptor-mediated mechanism 56
      - immunosuppression 162
      - angiogenesis
    - multiphase 3,137

- **Toxicokinetics** 4,320
  - absorption, uptake, distribution, excretion 531
  - bioaccumulation
  - metabolism 3,284
    - activation or deactivation 2,512
    - reactive oxygen species
  - toxicokinetic modeling 54

Figure 1. An overview of the CRAB3 user interface. By using the search query "benzo[*a*]pyrene," the CRAB3 tool retrieved all references containing information on benzo[*a*]pyrene from PubMed (accessed and retrieved in May 2020). The red oval-shaped marks are the distribution of literature according to the evidence they contain. The first column on the left, shows the classification of literature according to the scientific evidence (human/epidemiology study, animal study, *in vitro* cell experiments, and so on), the second column analyzes literature based on the evidence on carcinogenic MOAs (genotoxic and nongenotoxic MOAs categories and subcategories), and the last column shows literature with information on toxicokinetics. Note: AhR, aryl hydrocarbon receptor; CRAB3, Cancer Risk Assessment using Biomedical literature tool 3; MOA, mode of action; TNF, tumor necrosis factor.

B[*a*]P because the CRAB3 tool automatically considered literature on B[*a*]P while searching for pyrene. For other chemicals, such as anthracene, fluoranthene, and their related PAHs, we did a manual curation of CRAB3 data for anthracene and did not find major differences in the MOA categories (Excel Table S9). We believe that with manual curation, there will be minor-to-negligible changes on the proportions of MOA category, if any, so we did not proceed with it further. It should also be noted that other chemicals with similar parts of the name may have minor contributions to the outcome, if any. This is because one of the current limitations of CRAB3 is that it currently does not include entity grounding for chemicals, that is, it lacks the ability to recognize name variations of the same chemical and different chemicals with some similarity in names, such as pyrene and B[*a*]P.

A time–trend analysis of the literature concerning B[*a*]P was performed by grouping the CRAB3 data into three groups based on year of publication: 1970–2000, 2001–2010, and 2011–2019. Furthermore, to test whether CRAB3-generated MOA profiles were in line with IARC evaluations on evidence for key characteristics of carcinogens, a comparison with their evaluations was performed for B[*a*]P, dibenzo[*a,l*]pyrene (DB[*al*]P), benz[*a*]anthracene (B[*a*]A), dibenzo[*a,h*]anthracene (DB[*ah*]A, benzo[k] fluoranthene (B[*k*]F), diesel exhaust, coal tar, and crystalline silica (Krewski et al. 2019).

CHAT Analyses

The analysis of literature by CHAT is based on supervised machine learning to classify all of the >160 million sentences in the PubMed database according to the hallmarks of cancer taxonomy (Baker et al. 2017). Using co-occurrence–based association metrics, such as conditional probability (cprob), an association profile of co-occurrence for the search query term and the 37 cancer hallmarks categories (Table S2) (Hanahan and Weinberg 2000, 2011) was generated (Figure S3) that can be used for further comparison analyses with, for example, other chemicals. Cprob describes the strength of association for CHAT analyses, that is, the probability of a hallmark of cancer appearing in the literature given the occurrence of a chemical or a group of chemicals of interest.

For the current CHAT analyses, we used the same search terms as for CRAB3 for all the PAHs, PAH mixtures, and crystalline silica in the CHAT search function (March 2019) and used cprob as association metrics for co-occurrence analysis. Although CHAT was last updated in 2018 as compared with the CRAB3 tool, which is updated nightly, we believe that there were no major differences in the overall literature retrieved and analyzed both by the CRAB3 tool and CHAT. CHAT also lacks entity grounding, but manual curation was not possible with the feature of selecting and filtering of abstracts because this feature does not exist in CHAT.

Statistical Analysis

Principal component analysis (PCA) and hierarchical clustering analysis (HCA) were performed using SIMCA 16 software (Sartorius-Stedim Biotech). The PCA included the 22 priority PAHs with their respective MOA proportions. For each PAH, the MOA categories were included only where articles were found to contain the search term of interest. HCA calculated with Ward's method was performed on the obtained loading and score vectors. Dendrograms were prepared based on PC2 for the loadings and on PC1, -2, and -3 for the scores. The third component was added to increase the explanation of variance. In CRAB3 analysis, the tool used chi-squared test with Bonferroni correction to calculate statistical differences. The MOA profiles for different chemicals (individually or in groups) were compared, and statistically significant differences were computed using the chi-squared homogeneity test for each individual MOA category (positive vs. negative) and for each pair of chemicals (using a 2×2 contingency table). The individual p-values were then adjusted by a Bonferroni correction for the entire profile's p-values. A $p<0.05$ was considered significant for all statistical analyses.

Results

MOA Profiles and Cancer Hallmark Analyses for the 22 Priority PAHs

In total, the CRAB3 tool identified 53,944 references from PubMed (March 2019) concerning the 22 priority PAHs and classified 42% of abstracts/references (22,657 references of the total) as relevant for scientific evidence (including information on human and animal studies and on cell experiments) and 41% (22,031 references of the total) as relevant for carcinogenic MOAs (Table 2). An overview of the literature for individual PAHs is shown as Table S1. CRAB3-generated literature profiles for individual PAHs, over the selected carcinogenic MOAs, are shown in Figure 2A (Excel Table S2). Yellow-highlighted bars represent the MOA profile of B[*a*]P. The MOA profiles of individual PAHs show the literature heterogeneity of this complex group of chemicals. On average, and as expected, literature on genotoxic MOAs, including DNA adducts and mutations, represented the largest proportions (Figure 2A).

We extended our analyses on the 22 priority PAHs and evaluated PubMed literature by using CHAT. We evaluated the PubMed literature previously identified by the CRAB3 tool using CHAT to identify relevant evidence in the text to classify each reference according to the hallmarks of cancer taxonomy (Baker et al. 2017). As expected, the analyses showed genomic instability, namely, DNA damage, adducts, and mutations, as being the most common cancer hallmarks associated with the priority PAHs (Figure S4). Our data show that both CRAB3 and CHAT analyzed and identified genomic instability as the most commonly described effect for the priority PAHs.

MOA Profiles for Three Groups of PAHs Based on RPF Value

We tested whether CRAB3-generated MOA profiles could be useful in grouping priority PAH into subgroups according to the assigned RPF values (Collins et al. 1998; U.S. EPA 2010b). Grouping of PAHs and an overview of the literature for each group is shown in Table S1. CRAB3-generated MOA profiles for each group are shown in Figure 2B. A comparison of MOA profiles showed markedly higher proportions of literature in DNA adducts and mutations for PAHs in Group 3 as compared with PAHs in the two other groups (Figure 2B). In Group 1, oxidative stress and cell proliferation were among the most common MOAs, which is in line with their ability to rather function as tumor promoters than as genotoxic carcinogens (Misaki et al. 2016).

The ability of the CRAB3 tool to group PAHs based on literature analyses was further evaluated by PCA followed by HCA (see the "Material and Methods" section) and based on the MOA proportions shown in Figure 2. The first two principal components (PCs) explained 66.8% of the variance in the data. The loading vectors from the PCA did not show a clear separation between the genotoxic and nongenotoxic MOAs (Figure 3A). HCA identified three clusters of MOAs, of which two contained a mix of genotoxic and nongenotoxic MOAs (Figure 3B). For example, the MOAs including oxidative stress, epigenetics, strand breaks, adducts, aryl hydrocarbon receptor (AhR) activation, DNA repair, and transcriptional modification all belonged to the largest cluster, suggesting an overlap between the two

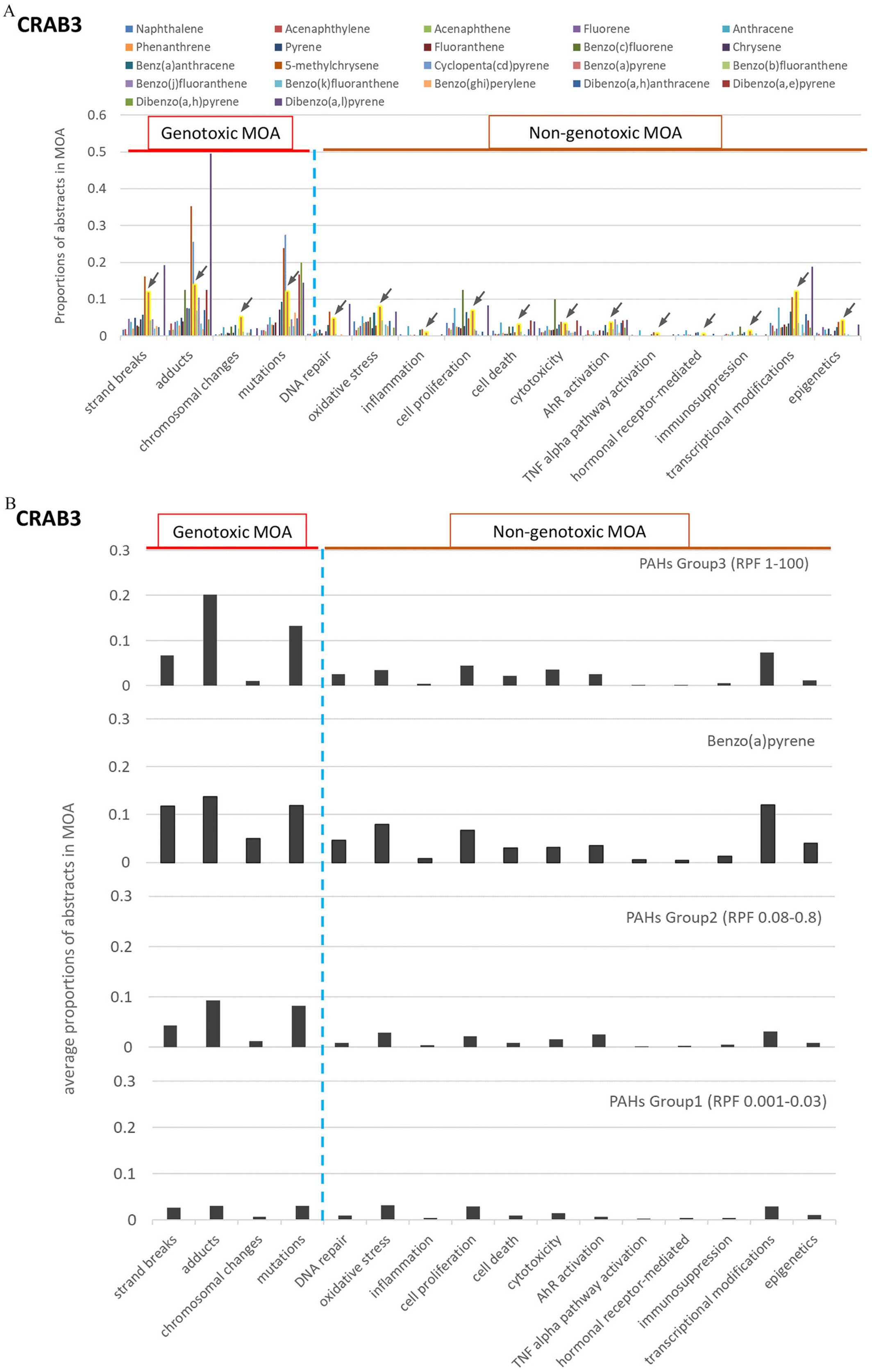


Figure 2. CRAB3 analysis of the literature for 22 priority PAHs. (A) CRAB3-generated mode of action (MOA) profiles of 22 polycyclic aromatic hydrocarbons (PAHs) identified under the U.S. EPA and European Union priority list (Jarvis et al. 2014). The MOA profile for reference compound benzo[*a*]pyrene (B[*a*]P) is shown as yellow-highlighted bars (pointed out with arrow). (B) CRAB3-generated MOA profiles of three groups of priority PAHs based on their relative potency factor (RPF) values and the reference compound B[*a*]P (Excel Tables S2 and S3). Data are presented as proportions of total references identified by CRAB3 in MOA categories, and the vertical dotted line separates the genotoxic and nongenotoxic MOA categories. Note: AhR, aryl hydrocarbon receptor; CRAB3, Cancer Risk Assessment using Biomedical literature tool 3; EPA, Environmental Protection Agency; TNF, tumor necrosis factor.

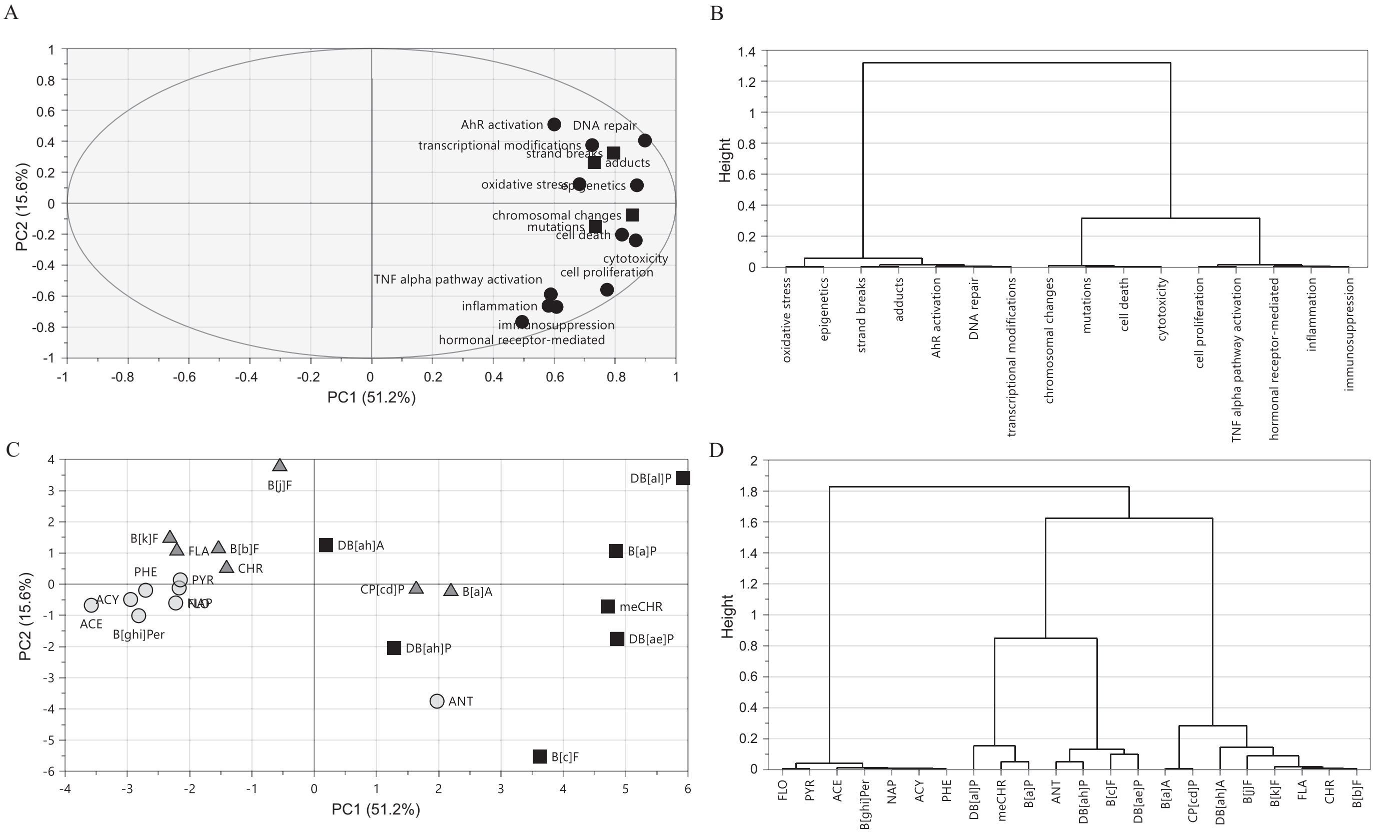


Figure 3. Principal component and hierarchical clustering analyses of CRAB3-generated mode of action (MOA) profiles of PAHs. (A) Scatter plot of loading vectors of the 16 MOAs shown in Figure 2 (Excel Table S3) in the first and second principal components (PCs). Genotoxic MOAs are shown as boxes and nongenotoxic MOAs as circles. (B) Dendrogram of the loadings in PC2. (C) Scatter plot of the score vectors of the 22 PAHs included in the CRAB3 analysis in the first and second PCs. Abbreviations are explained in Table 1. PAHs with low, moderate, and high RPF values are shown as light gray circles, gray triangles, and black boxes, respectively. (D) Dendrogram of the scores in the three first PCs. Note: ACE, acenaphthene; ACY, acenaphthylene; AhR, arylhydrocarbon receptor; ANT, anthracene; B[a]A, benz[a]anthracene; B[a]P, benzo[a]pyrene; B[b]F, benzo[b]fluoranthene; B[c]F, benzo[c]fluorene; B[ghi]Per, benzo[ghi]perylene; B[j]F], benzo[j]fluoranthene; B[k]F, benzo[k]fluoranthene; CHR, chrysene; CP[cd]P, cyclopenta[cd]pyrene; CRAB3, Cancer Risk Assessment using Biomedical literature tool3; DB[ae]P, dibenzo[a,e]pyrene; DB[ah]A, dibenzo[a,h]anthracene; DB[ah]P, dibenzo[a,h]pyrene; DB[al]P, dibenzo[a,l]pyrene; FLA, fluoranthene; FLO, fluorene; meCHR, 5-methylchrysene; NAP, naphthalene; PAH, polycyclic aromatic hydrocarbon; PHE, phenanthrene; PYR, Pyrene; RPF, relative potency factor; TNF, tumor necrosis factor.

overarching types of MOAs. The only nongenotoxic MOA cluster included cell proliferation, tumor necrosis factor-alpha (TNFα) pathway activation, hormonal receptor-mediated, inflammation, and immunosuppression.

The score plot of the two first components suggested a separation between low (0.001–0.03) and moderate (0.08–0.8) from high (1–100) RPF PAHs in PC1, although with some overlap (Figure 3C). Similarly, the low RPF PAHs were separated from the moderate in PC2, together indicating that the CRAB3 tool was able to cluster PAHs according to their carcinogenic potency. This clustering was further supported by HCA, which suggested three clusters with high dissimilarity (Figure 3D). One cluster included all the low RPF PAHs except anthracene (ANT). Similarly, one cluster included all the moderate RPF PAHs and DB[*ah*]A, which had a wide range of published RPF values spanning from moderate to high potency (0.1–10) (Table 1). The final cluster included the majority of the high RPF PAHs plus ANT, which had an RPF range of 0–0.1 (Table 1). The MOAs with the largest contribution to the score of most high RPF PAHs were a mix of genotoxic and nongenotoxic MOAs, whereas for benzo[*c*]fluorene and ANT, nongenotoxic MOAs inflammation and the TNFα pathway activation were the largest contributors. Together, the data from the PCA and HCA suggest that the CRAB3 tool has the capacity to evaluate literature concerning the complex group of PAHs and group them according to their relative potencies.

Cancer Hallmark Analyses of Three Groups of PAHs Based on RPF Value

We further analyzed the literature concerning three groups of priority PAHs, by using CHAT. The data from CHAT analyses showed that genomic instability, namely, DNA adducts, damage, and mutations, was the most common cancer hallmark associated with PAH Group 3 (RPF > 1) (Figure 4A). In contrast, tumor-promoting inflammation, sustaining proliferating signaling, and evading growth suppressor were more often reported with PAH Groups 1 and 2 (RPF < 1) (Figure 4B,C). CHAT data thus supports that PAHs with an RPF > 1 act via genotoxic MOAs, that is, through adducts and mutations to induce cancer, and PAHs with an RPF < 1 have tumor-promoting properties, acting via nongenotoxic MOAs, that is, through inflammatory responses, oxidative stress, and cell proliferation to promote cancer (Misaki et al. 2016). Thus, CHAT data are in line with CRAB3. Furthermore, the data also suggest that interactions between high and low potency PAHs may occur, in a combined exposure setting.

Comparing MOA Profile for B[a]P with Dibenzo[a,l]pyrene

DB[*al*]P, also called dibenzo[*def,p*]chrysene, is one of the most potent PAHs, with RPF values ranging up to 100 (Luch 2009). DB[*al*]P is classified under Group 2A (probable human carcinogen) by the IARC (2016) but is very rarely included in

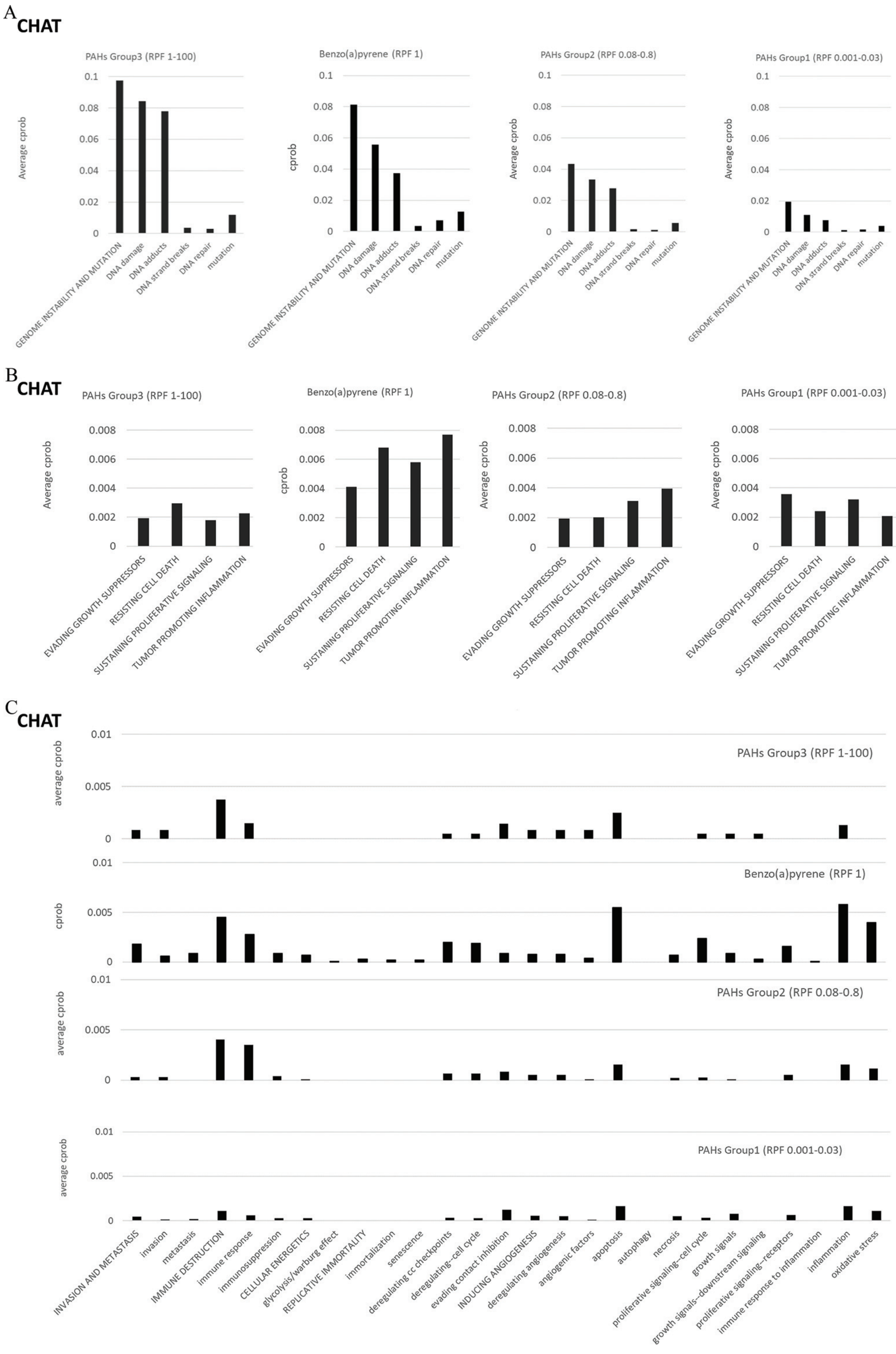


Figure 4. CHAT tool analyses of the literature concerning three groups of priority PAHs based on their relative potency factor (RPF) values. (A) Comparison of a CHAT-generated literature profile on genomic instability and mutation for the three groups of priority PAHs based on their RPF values and the reference compound benzo[*a*]pyrene (B[*a*]P); (B) comparison of CHAT-generated literature profile on selected cancer hallmarks, including evading growth suppressors, resisting cell death, sustaining proliferative signaling and tumor-promoting inflammation, for the three groups of priority PAHs based on their RPF values and the reference compound B[*a*]P; and (C) CHAT-generated literature profiles on all other selected cancer hallmarks and associated processes for the three groups of priority PAHs and comparison with B[*a*]P. Summary data are shown in the supplemental Excel file Table S4. Data are presented as conditional probability (cprob) showing the probability of a hallmark of cancer appearing in the literature given the occurrence of a particular chemical or a group of chemicals of interest. Note: cc, cell cycle; CHAT, Cancer Hallmarks Analytics Tool; PAH, polycyclic aromatic hydrocarbon.

environmental monitoring or human health risk assessment (Andersson and Achten 2015). We analyzed the literature to test whether the difference in the relative potencies between DB[*al*]P and B[*a*]P can be identified by the CRAB3 tool and CHAT. CRAB3 analyses of the amount of literature on scientific evidence for carcinogenicity shows that, compared with B[*a*]P, DB[*al*]P is much less studied (Figure 5A). The total number of PubMed references for DB[*al*]P was 228, compared with 11,650

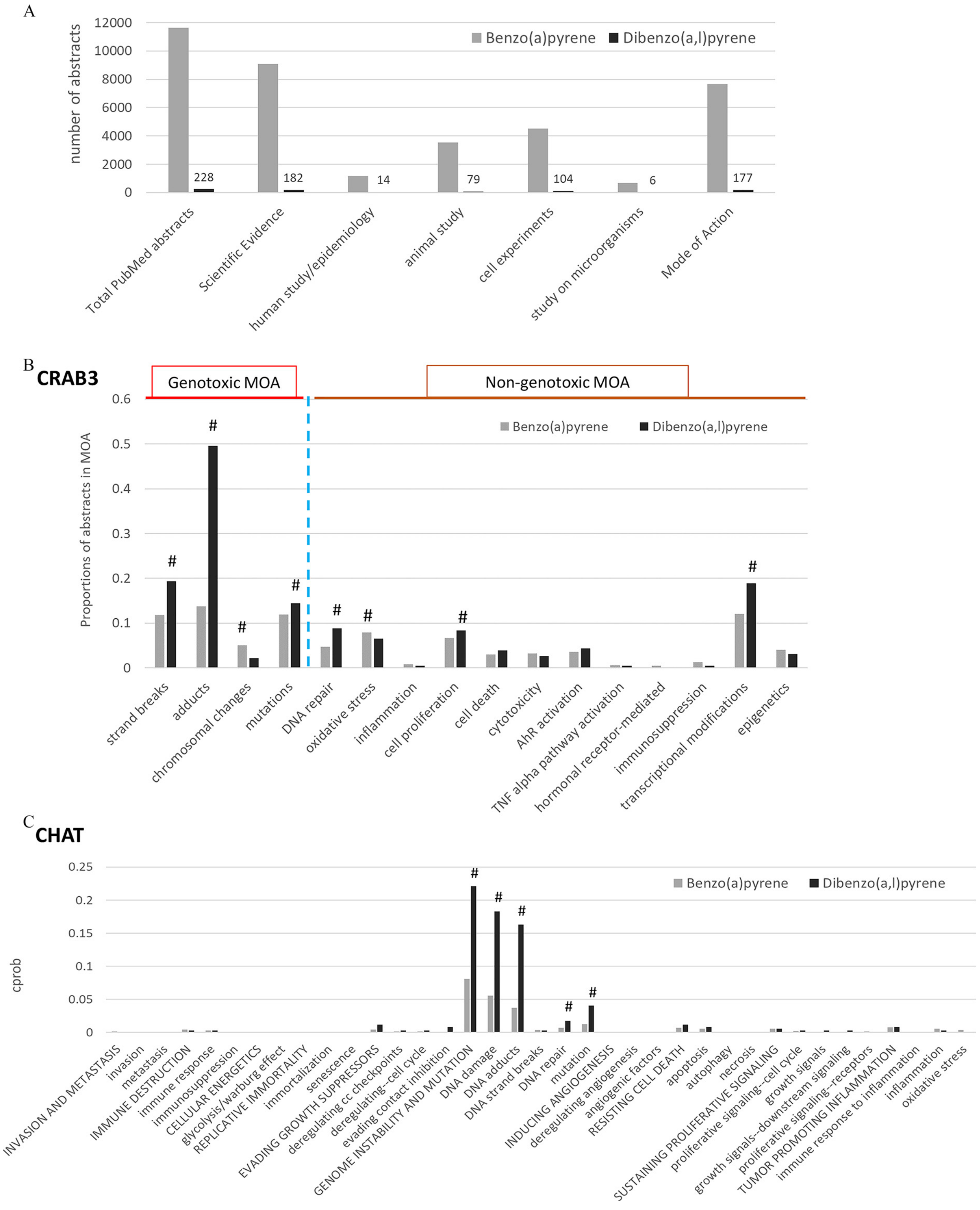


Figure 5. Comparison of the literature profile between benzo[*a*]pyrene (B[*a*]P) and dibenzo[*a,l*]pyrene (DB[*al*]P). (A) An overview of total PubMed literature and the literature identified as scientific evidence and as MOA by the CRAB3 tool concerning B[*a*]P and DB[*al*]P; (B) CRAB3-generated MOA profiles for B[*a*]P and the most potent PAH, DB[*al*]P; and (C) CHAT-generated literature profile on cancer hallmarks for B[*a*]P and DB[*al*]P. Data are presented as proportions of total literature in MOA categories for CRAB3 analyses (a vertical dotted line separates the genotoxic and nongenotoxic MOA categories) and conditional probability (cprob) as a strength of association for CHAT analyses (Excel Table S5), showing the probability of a hallmark of cancer appearing in literature given the occurrence of a particular chemical or a group of chemicals of interest. Statistically significant differences were computed using the chi-squared homogeneity test for each individual MOA category (positive vs. negative) and for each pair of chemicals (using a 2 × 2 contingency table). The individual *p*-values were then adjusted by a Bonferroni correction for the entire profile's *p*-values. $p < 0.05$ is considered significant". #, $p < 0.05$ significantly different from each other. Note: AhR, aryl hydrocarbon receptor; cc, cell cycle; CHAT, Cancer Hallmarks Analytics Tool; CRAB3, Cancer Risk Assessment using Biomedical literature tool 3; MOA, mode of action; TNF, tumor necrosis factor.

references for B[*a*]P. However, CRAB3-generated MOA profiles showed that significantly higher proportions of the literature ($p > 0.01$) on adducts, strand breaks, mutations, DNA repair, and cell proliferation are associated with DB[*al*]P as compared with B[*a*]P (Figure 5B).

In addition, we performed time–trend analysis on the large volume of literature concerning B[*a*]P (Figure S5). The analysis demonstrated that research during 1970–2000 was more focused on investigating genotoxic MOAs, whereas more recently, in 2001–2010 and 2011–2019, more emphasis has been on nongenotoxic MOAs. This suggests that the nongenotoxic MOAs of B[*a*]P is a relatively neglected research area.

Furthermore, literature analyses by CHAT on cancer hallmarks and associated cellular processes (Figure 5C) showed that there was significantly higher probability ($p < 0.01$) of the hallmark of cancer genomic instability and mutations appearing in the literature on DB[*al*]P as compared with B[*a*]P. CHAT analysis also highlights other differences with mechanistic details, for example, in evading contact inhibition between these compounds. Deregulation of contact inhibition has been proposed for nongenotoxic AhR ligands including PAHs; however, it is not known whether genotoxic ligands may have similar effects (Dietrich and Kaina 2010). Manual evaluation of the evading contact inhibition node showed that DB[*al*]P induced epidermal hyperplasia that may promote tumor development *in vivo* (Casale et al. 2000). Our evaluation by CHAT suggesting that evading contact inhibition may be activated by genotoxic PAHs indicated that the data is lacking in this area and need to be investigated further. In summary, the results from these analyses support and confirm the importance of including DB[*al*]P in human health risk assessments.

Comparing MOA Profiles for Chrysene to 5-Methylchrysene and Pyrene to 1-Nitropyrene

5-Methylchrysene is reasonably anticipated to be a human carcinogen owing to its higher carcinogenic potential as compared with chrysene (NTP 2016). We evaluated literature on chrysene and 5-methylchrysene (Figure 6A,B) to explore whether the differences in carcinogenic potency due to the presence of the methyl group can be supported by CRAB3 and CHAT analysis. A comparison of literature profiles (Figure 6A,B) shows that 5-methylchrysene had significantly higher proportions of literature on adducts, strand breaks, and mutations as compared with chrysene ($p < 0.01$), which is in line with the current understanding about the carcinogenic potential of 5-methylchrysene as compared with chrysene.

Pyrene has been classified by the IARC under Group 3, meaning that the evidence of carcinogenicity is inadequate in humans and inadequate or limited in experimental animals (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans 2010). We analyzed the literature on pyrene and compared it with its nitro-substituted derivative, 1-nitropyrene, which is classified by the IARC under Group 2A as a probable human carcinogen (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans 2010). A comparison of MOA profiles (Figure 6C,D) shows that, compared with pyrene, there was a significantly higher probability of genotoxic MOAs, including mutations, adducts, and strand breaks, as well as nongenotoxic MOAs, including DNA repair, oxidative stress, cell proliferation, and epigenetics, appearing in the 1-nitropyrene literature. These data are consistent with findings suggesting that 1-nitropyrene induces DNA damage and oxidative stress and suppresses DNA repair in rats (Li et al. 2017).

Structural alterations, namely, substituting methyl or nitro groups to PAHs, may increase the carcinogenic potential of the parent compound, as is the case for 5-methylchrysene and 1-nitropyrene. Our data indicate that both CRAB3 and CHAT analyses offered additional evidence that informs how these chemicals could interact in mixture.

Evaluation of B[a]P as a Marker for PAHs in Complex Air Emissions Using CRAB3 and CHAT

In the European Union, B[*a*]P is used as a surrogate marker for PAHs in air. However, both the World Health Organization (WHO) and the European Food Safety Agency (EFSA) have concluded that the surrogate marker approach is likely to misestimate the actual risk given that co-occurring substances are also carcinogenic (EFSA 2008; WHO 2000). The representativeness of B[*a*]P for complex mixtures were tested by analyzing literature (3,554 references) for air emissions containing complex PAH mixtures, including diesel exhaust, and coke oven emission, and the PAH mixture, coal tar, by using CRAB3 and CHAT analysis, followed by a comparison with B[*a*]P.

Diesel exhaust is the most well-studied PAH mixture, with more than 2,000 references included in the analyses (Figure 7A). CRAB3- and CHAT-generated data showed that DNA strand breaks, adducts, and chromosomal changes were significantly more likely ($p < 0.05$) to be associated with coke oven emission literature, whereas oxidative stress, tumor-promoting inflammation, and immunosuppression were the most common carcinogenic MOAs associated with diesel exhaust as compared with B[*a*]P (Figure 7B,C). Although the MOA profile of coal tar shows mutations, strand breaks, and adducts as the most common MOAs, the literature proportions were much lower than those for B[*a*]P. Furthermore, detailed analyses by CHAT (Figure 7D), showed that immune response and evading contact inhibition were more commonly associated with diesel exhaust and coal tar literature, respectively, as compared with B[*a*]P. Considering the individual components of each PAH mixture, which also includes B[*a*]P, our evaluations suggest that the combined carcinogenic potential of all PAHs in a mixture may be different (likely due to additive or interactive effects) than what is estimated by using B[*a*]P as a marker.

Comparison of CRAB3 vs. IARC Evaluation of Key Characteristics

Furthermore, to test whether CRAB3-generated MOA profiles are in line with the IARC evaluations on the evidence for key characteristics of carcinogens (Krewski et al. 2019), we did a comparison of their evaluations for several PAHs, including B[*a*]P, DB[*al*]P, B[*a*]A, DB[*ah*]A, B[*k*]F, diesel exhaust, coal tar, and crystalline silica (Table S3). Our evaluation showed that CRAB3 analysis correctly predicted the evidence on key characteristics as identified by the IARC regarding these PAHs. For example, as shown in Table S3, seven key characteristics were highlighted by the IARC in the evaluation of B[*a*]P (with evidence from one or more human, animal, or *in vitro* sources), and CRAB3 analysis correctly identified the evidence on all of the key characteristics, namely, metabolic activation, genotoxicity, oxidative stress, altered cell proliferation/cell death, epigenetic alterations, immunosuppression, and receptor-mediated effects. The agreement between CRAB3 and IARC evaluations adds further strength to the applicability of our TM approach in hazard identification.

Identification of Mechanisms of Potential Importance for Interactions between Crystalline Silica and PAHs (B[a]P) Using CRAB3 and CHAT

Exposure to airborne particles such as crystalline silica is a major global environmental and occupational health hazard (Leung et al. 2012). It has been hypothesized that simultaneous exposure to

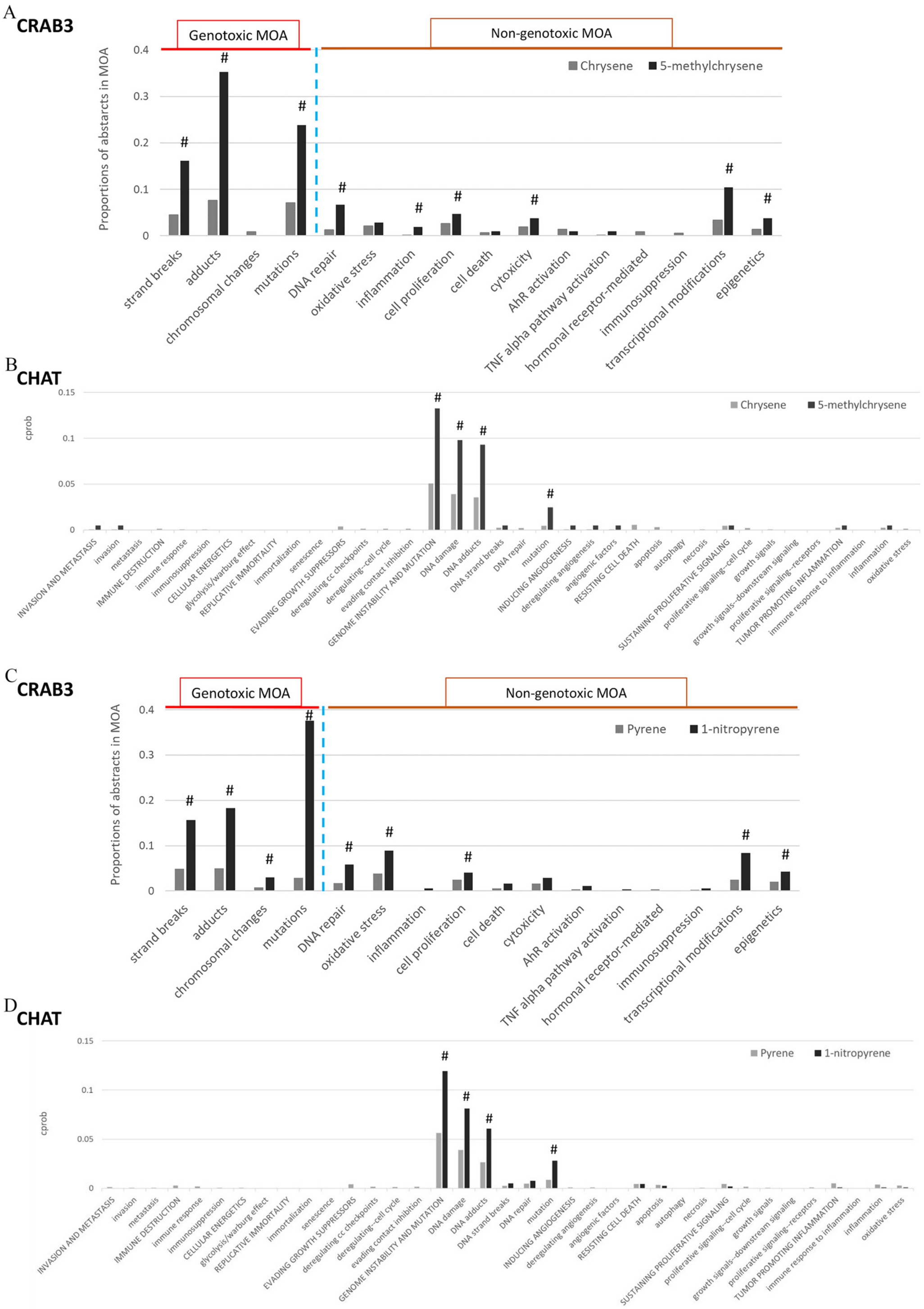


Figure 6. Substitution of methyl or nitro group and the carcinogenic MOA of PAHs. (A,B) CRAB3- and CHAT-generated literature profiles of chrysene and 5-methylchrysene on carcinogenic MOA and cancer hallmarks, respectively. (C,D) CRAB3- and CHAT-generated literature profiles of pyrene and 1-nitropyrene on carcinogenic MOA and cancer hallmarks, respectively. Data are shown as proportions of total literature in MOA categories for CRAB3 analyses (a vertical dotted line separates the genotoxic and nongenotoxic MOA categories) and conditional probability (cprob) as a strength of association for CHAT analyses (Excel Table S6), showing the probability of a hallmark of cancer appearing in literature given the occurrence of a particular chemical or a group of chemicals of interest. Statistically significant differences were computed using the chi-squared homogeneity test for each individual MOA category (positive vs. negative) and for each pair of chemicals (using a 2×2 contingency table). The individual p-values were then adjusted by a Bonferroni correction for the entire profile's p-values. $p < 0.05$ is considered significant". #, $p < 0.05$ significantly different from each other. Note: AhR, aryl hydrocarbon receptor; cc, cell cycle; CHAT, Cancer Hallmarks Analytics Tool; CRAB3, Cancer Risk Assessment using Biomedical literature tool 3; MOA, mode of action; PAH, polycyclic aromatic hydrocarbon; TNF, tumor necrosis factor.

known carcinogens, such as B[*a*]P, can possibly increase the carcinogenicity of crystalline silica (Downward et al. 2014; Vermeulen et al. 2011). Therefore, we analyzed literature (935 references) on crystalline silica and compared the results with those for B[*a*]P. Both CRAB3- and CHAT-generated literature profiles showed tumor-promoting inflammation, immunosuppression, and oxidative

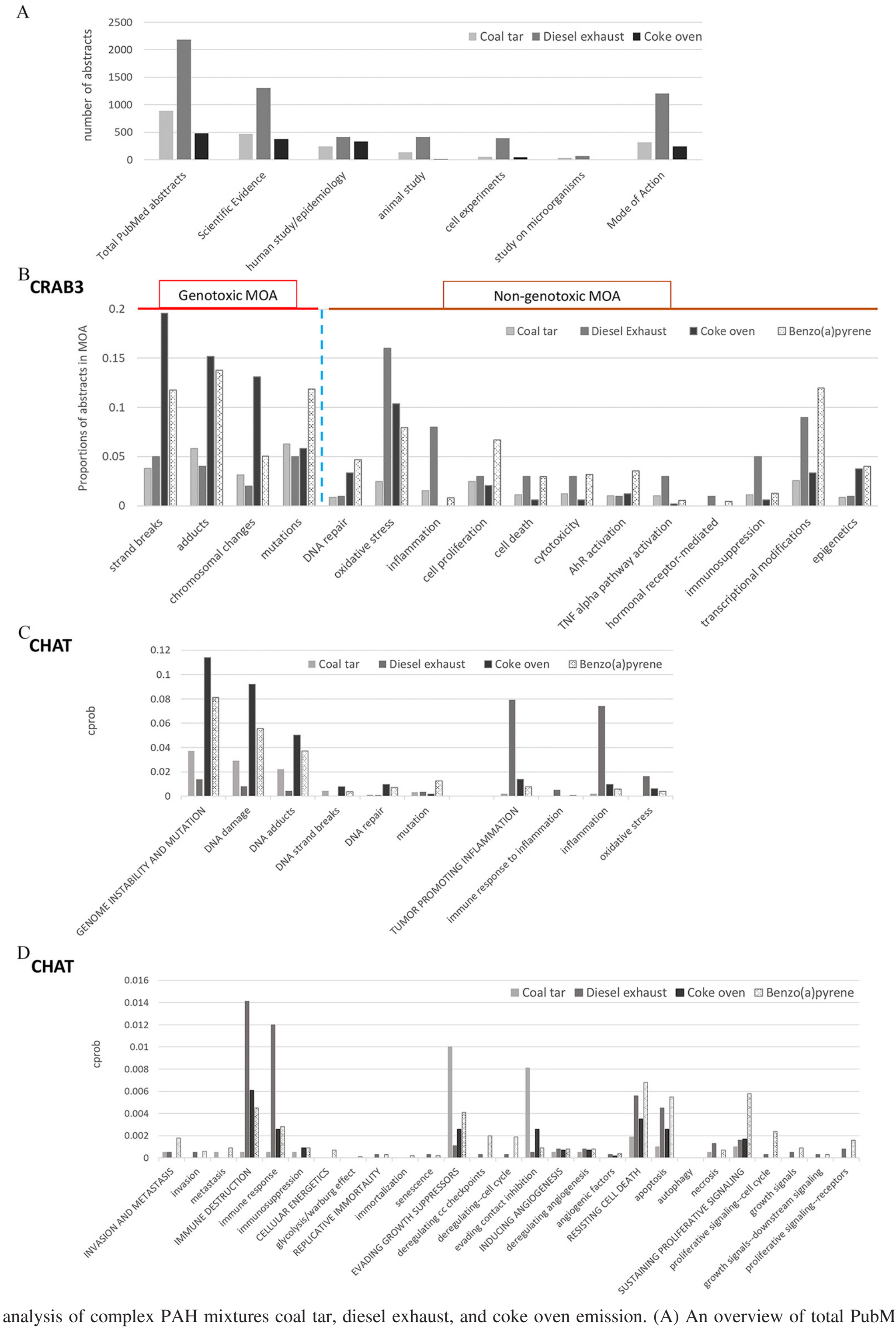


Figure 7. Literature analysis of complex PAH mixtures coal tar, diesel exhaust, and coke oven emission. (A) An overview of total PubMed literature and the literature identified by the CRAB3 tool as scientific evidence and relevant for MOA, for coal tar, diesel exhaust, and coke oven emission; (B) CRAB3-generated MOA profiles concerning coal tar, diesel exhaust, and coke oven emission and comparison with the reference compound benzo[*a*]pyrene (B[*a*]P); and (C, D) CHAT-generated literature profile on cancer hallmarks concerning coal tar, diesel exhaust, and coke oven emission and comparison with the reference compound B[*a*]P. Data are shown as proportions of total literature in MOA categories for CRAB3 analyses (a vertical dotted line separates the genotoxic and non-genotoxic MOA categories) and conditional probability (cprob) as a strength of association for CHAT analyses (Excel Table S7), showing the probability of a hallmark of cancer appearing in literature given the occurrence of a particular chemical or a group of chemicals of interest. Statistically significant differences are computed using the chi-squared homogeneity test for each individual MOA category (positive vs. negative) and for each pair of chemicals (using a 2 × 2 contingency table). The individual *p*-values were then adjusted by a Bonferroni correction for the entire profile's *p*-values. $p < 0.05$ is considered significant. Note: AhR, aryl hydrocarbon receptor; cc, cell cycle; CHAT, Cancer Hallmarks Analytics Tool; CRAB3, Cancer Risk Assessment using Biomedical literature tool 3; MOA, mode of action; PAH, polycyclic aromatic hydrocarbon, TNF, tumor necrosis factor.

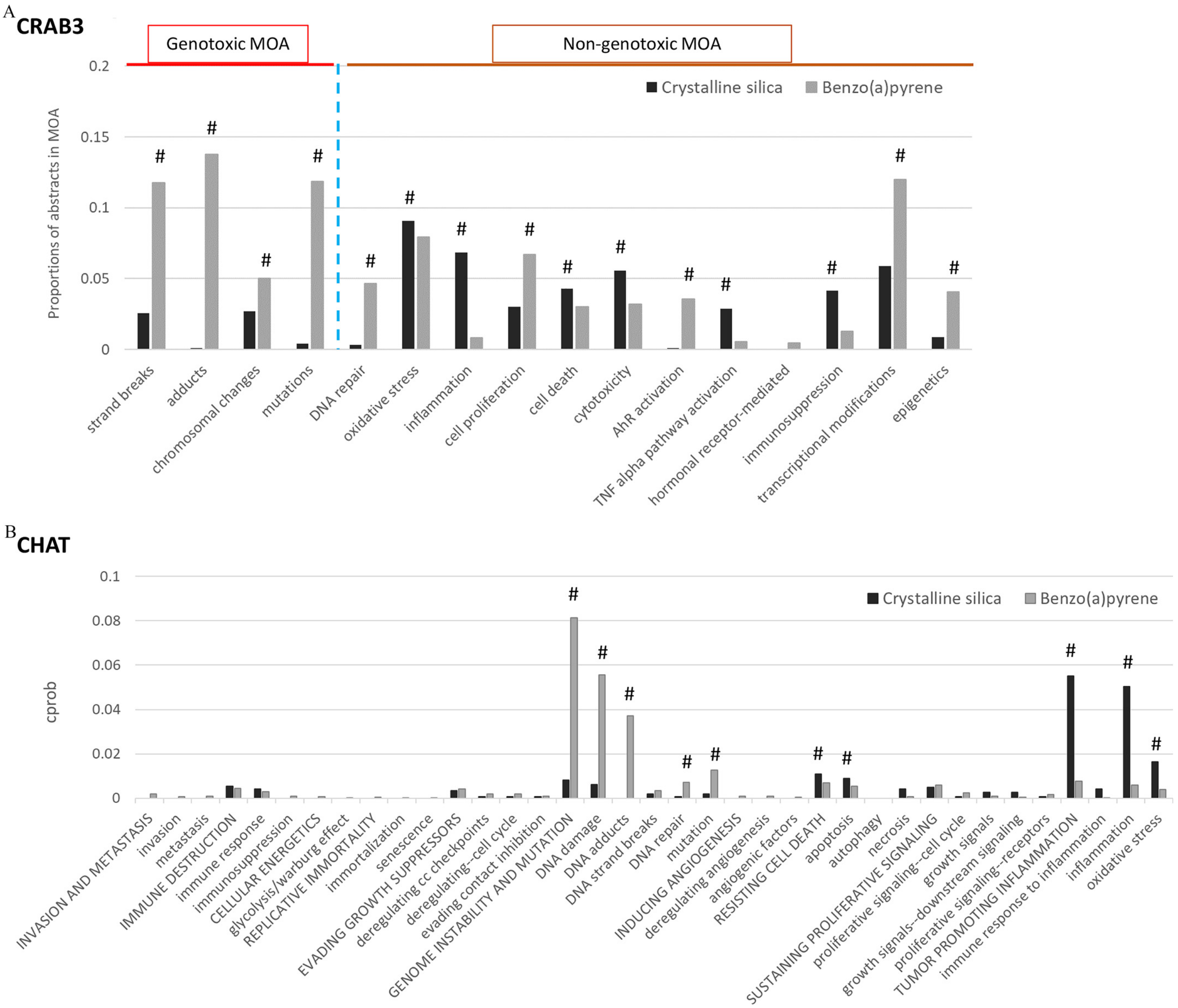


Figure 8. Literature analyses of crystalline silica and benzo[*a*]pyrene (B[*a*]P). (A) CRAB3-generated literature profiles of crystalline silica and comparison with the reference PAH B[*a*]P. (B) CHAT analyses of the literature on cancer hallmarks and associated processes, concerning crystalline silica and comparison with B[*a*]P. Data are shown as proportions of total literature in the MOA categories for CRAB3 (a vertical dotted line separates the genotoxic and nongenotoxic MOA categories) and conditional probability (cprob) as a strength of association for CHAT analyses (Excel Table S8), showing the probability of a hallmark of cancer appearing in literature given the occurrence of a particular chemical or a group of chemicals of interest. Statistically significant differences are computed using the chi-squared homogeneity test for each individual MOA category (positive vs. negative) and for each pair of chemicals (using a 2×2 contingency table). The individual p-values were then adjusted by a Bonferroni correction for the entire profile's p-values. $p < 0.05$ is considered significant. #, $p < 0.05$ significantly different from each other. Note: AhR, aryl hydrocarbon receptor; cc, cell cycle; CHAT, Cancer Hallmarks Analytics Tool; CRAB3, Cancer Risk Assessment using Biomedical literature tool 3; MOA, mode of action; PAH, polycyclic aromatic hydrocarbon; TNF, tumor necrosis factor.

stress as the most common outputs, with a significantly higher possibility of being associated with the crystalline silica literature than with the B[*a*]P literature ($p < 0.01$) (Figure 8A,B). These findings are in line with the current understanding about crystalline silica (Satpathy et al. 2015). Our analyses indirectly support the notion that, in a combined exposure to crystalline silica and B[*a*]P, crystalline silica may lead to an enhanced carcinogenic potential of B[*a*]P.

Discussion

By using a combination of our TM tools CRAB3 and CHAT, we present a case study of PAHs based on the PubMed literature (March 2019) and show that our tools can identify similarities and differences in carcinogenic MOAs. First, we analyzed the literature concerning the 22 priority PAHs and tested whether the information generated by the CRAB3 tool can be useful to classify PAHs into subgroups. To evaluate the classification, we employed RPF values assigned to these PAHs and found a correlation between CRAB3- and CHAT-generated profiles and RPFs and that the CRAB3 tool could group PAHs according to their potency. We also found that differences in the carcinogenic potential of PAHs and PAH derivatives, namely, methylated- or nitrated-PAHs, and of other chemicals, for example, crystalline silica, can be identified. An unexpected finding was that the tools revealed major mechanistic differences in the characterization of naturally occurring PAH mixtures. These data show that CRAB3 and CHAT analyses gave similar answers to our questions and had the capacity to evaluate a large amount of literature concerning the complex group of PAHs. Moreover, comparison of CRAB3 analyses with the IARC evaluation in identifying key characteristics on carcinogens showed similar evidence and correlated with traditionally conducted assessments.

The classification of literature by each of these tools is based on different taxonomies and TM techniques. Nevertheless, the output data from CRAB3 and CHAT complemented each other in this study. CRAB3 classification of literature is based on the evidence for carcinogenic MOAs, and the tool was developed primarily for cancer risk assessment purposes (Korhonen et al. 2012). CHAT analysis was developed primarily for cancer researchers and for classifying cancer literature according to cancer hallmarks and associated processes (Baker et al. 2017). CHAT analysis has the capacity to identify more detailed end points not covered by the CRAB taxonomy. Because CHAT analysis is built on concepts employed in basic research, using CHAT in combination with the CRAB3 tool might be particularly helpful for basic researchers interested in risk assessment issues.

Our data suggest that both TM tools can identify literature profiles that reflect potency and can, for example, support grouping chemicals with similar relative potency. Comparing MOA profiles for structurally related single chemicals within the PAH group and with strongly differing potency, such as chrysene and 5-methylchrysene, complemented the analysis based on RPF values by showing that both tools could differentiate literature on structurally similar chemicals but with widely differing potency. Even analysis of B[*a*]P and DB[*al*]P, which are structurally related but have largely differing sizes of literature bases, revealed expected results with selectively more data on DNA damage-related effects for the latter. These results were based on PAHs, for which RPFs previously have been developed, but suggest that these tools also could be used for potency grouping of chemicals that lack potency factors.

Cancer risk assessment of a complex group of chemicals, such as PAHs, is challenging because exposure to PAHs occurs mainly in mixtures. Diesel exhaust, coal tar, and coke oven emission are among the most well-studied environmental PAH mixtures, and thus large volumes of literature are available for TM-based analyses. These three natural mixtures of PAHs exhibit quite different literature profiles. For example, the MOA profile of coke oven emission shows a higher than expected co-occurrence with strand breaks and chromosomal changes in the text as compared with B[*a*]P. Similarly, data on diesel exhaust show a higher than expected co-occurrence with oxidative stress, inflammation, and immunosuppression in the text as compared with B[*a*]P. Although all these mixtures are classified as human carcinogens (IARC 2016; IARC Working Group on the Evaluation of Carcinogenic Risks to Humans 2010), it is the levels of only B[*a*]P alone that are used for controlling these emissions and assuring air quality in relation to cancer risk for the general population within the European Union. In addition, there are diesel emission standards managed by, for example, the European Union and the U.S. EPA for particulate matter, nitrogen oxides, and carbon monoxide to limit noncancer health effects (Khalek et al. 2015). The progression of these emission standards has resulted in the development of new-technology diesel engines with reduced emission levels, including PAHs (Khalek et al. 2015). In addition, in 2019, the European Union amended their directive on the protection of workers from exposure to carcinogens to also include diesel exhaust, with a limit value based on elemental carbon (EU 2019). Our findings suggest that we may need to reevaluate the use of B[*a*]P as a marker for the carcinogenic risk of PAH mixtures and warrant development of new strategies, as has been discussed recently (Andersson and Achten 2015; Jarvis et al. 2014).

In addition, the possibility that PAHs and crystalline silica may interact was suggested by CRAB3 and CHAT analysis, as have been proposed in previous studies (Downward et al. 2014; Vermeulen et al. 2011). The toxicological profiles created by the CRAB3 tool and CHAT respectively for crystalline silica were similar, once again documenting that these two tools can be used for confirming robustness for carcinogens other than PAHs. By employing the CHAT tool and comparing the cancer hallmarks profiles of B[*a*]P and crystalline silica, we hypothesize that in a co-exposure setting, these chemicals could cause carcinogenic synergies. They may act in accordance with experimental studies employing initiation–promotion protocols (Cohen and Ellwein 1990; Solano et al. 2016). They may also act through different and, presumably, complementing cancer hallmarks, as has been evaluated in the Halifax project (Goodson et al. 2015).

Both CRAB3 and CHAT analysis have many advantages when evaluating large volumes of textual data in abstracts; however, they have some limitations. One of the major limitations of these tools is that they do not have the capacity to distinguish between positive and negative findings or, for example, between down- vs. up-regulation of inflammatory effects. The preferences in the types of research performed, that is, the extent at which a chemical has been studied, may also have an influence on the output. However, it should be noted that the output from these tools is based on available literature and the lack of literature does not necessarily mean the lack of association of a particular MOA or hallmark with the chemical. Furthermore, methodological differences in the study design, that is, dose, route of exposure, and so on, need to be examined manually. Future development of these tools includes features such as identification of positive vs. negative data, consideration of study design, and bibliometric indicators. We plan to upgrade the tools such that chemicals identified by the CRAB3 tool will be assigned a Chemical Abstracts Service number. For classification of compounds with limited amount of human or animal toxicity data, we also plan to integrate structure–activity relationship modeling into our TM tools, as suggested in our previous study (Papamokos and Silins 2016).

Conclusions

In summary, we show that our TM tools CRAB3 and CHAT can be used for literature review in cancer risk assessment of groups of chemicals. Using 22 priority PAHs and complex PAH mixtures, we observed that the similarities and differences between the PAHs identified by our tools can be useful for grouping these chemicals based on their RPF values. Moreover, these tools can facilitate the identification of possible interactions within a group of chemicals such as PAHs and between complex mixtures. A surprising finding was that naturally occurring mixtures that include PAHs exhibited different literature profiles. This case study using PAHs suggests that our tools can be useful not only for risk assessors handling groups of similar chemicals and mixtures but possibly also for cancer researchers addressing cancer risk assessment issues, including gaps in knowledge.

We found that these tools rapidly give literature overviews that complement and enrich the outputs. Used together, they can effectively assist manual reading of the articles. Not only risk assessors, but researchers can also benefit when addressing risk assessment issues. By using these tools, researchers can get a better understanding of how cancer risk assessment might gain from using new and cutting-edge science.

Acknowledgments

This research was supported by the Medical Research Council-UK and Swedish Research Council for Health, Working Life and Welfare.

References

Ali I, Guo Y, Silins I, Högberg J, Stenius U, Korhonen A. 2016. Grouping chemicals for health risk assessment: a text mining-based case study of polychlorinated biphenyls (PCBs). Toxicol Lett 241:32–37, PMID: 26562772, https://doi.org/10.1016/j.toxlet.2015.11.003.

Andersson JT, Achten C. 2015. Time to say goodbye to the 16 EPA PAHs? Toward an up-to-date use of PACs for environmental purposes. Polycycl Aromat Compd 35(2–4):330–354, PMID: 26823645, https://doi.org/10.1080/10406638.2014.991042.

Asada S, Sasaki K, Tanaka N, Takeda K, Hayashi M, Umeda M. 2005. Detection of initiating as well as promoting activity of chemicals by a novel cell transformation assay using v-Ha-*ras*-transfected BALB/c 3T3 cells (Bhas 42 cells). Mutat Res 588(1):7–21, PMID: 16260176, https://doi.org/10.1016/j.mrgentox.2005.07.011.

Baan R, Grosse Y, Straif K, Secretan B, El Ghissassi F, Bouvard V, et al. 2009. A review of human carcinogens—part F: chemical agents and related occupations. Lancet Oncol 10(12):1143–1144, PMID: 19998521, https://doi.org/10.1016/S1470-2045(09)70358-4.

Baker S, Ali I, Silins I, Pyysalo S, Guo Y, Högberg J, et al. 2017. Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer. Bioinformatics 33(24):3973–3981, PMID: 29036271, https://doi.org/10.1093/bioinformatics/btx454.

Casale GP, Cheng Z, Liu Jn, Cavalieri EL, Singhal M. 2000. Profiles of cytokine mRNAs in the skin and lymph nodes of SENCAR mice treated epicutaneously with dibenzo[*a,l*]pyrene or dimethylbenz[*a*]anthracene reveal a direct correlation between carcinogen-induced contact hypersensitivity and epidermal hyperplasia. Mol Carcinogen 27(2):125–140, PMID: 10657905, https://doi.org/10.1002/(SICI)1098-2744(200002)27:2<125::AID-MC8>3.0.CO;2-0.

Cohen SM, Ellwein LB. 1990. Cell proliferation in carcinogenesis. Science 249(4972):1007–1011, PMID: 2204108, https://doi.org/10.1126/science.2204108.

Collins JF, Brown JP, Alexeeff GV, Salmon AG. 1998. Potency equivalency factors for some polycyclic aromatic hydrocarbons and polycyclic aromatic hydrocarbon derivatives. Regul Toxicol Pharmacol 28(1):45–54, PMID: 9784432, https://doi.org/10.1006/rtph.1998.1235.

Dietrich C, Kaina B. 2010. The aryl hydrocarbon receptor (AhR) in the regulation of cell–cell contact and tumor growth. Carcinogenesis 31(8):1319–1328, PMID: 20106901, https://doi.org/10.1093/carcin/bgq028.

Downward GS, Hu W, Large D, Veld H, Xu J, Reiss B, et al. 2014. Heterogeneity in coal composition and implications for lung cancer risk in Xuanwei and Fuyuan counties, China. Environ Int 68:94–104, PMID: 24721117, https://doi.org/10.1016/j.envint.2014.03.019.

EC (European Commission). 2002. *Opinion of the Scientific Committee on Food on the risks to human health of Polycyclic Aromatic Hydrocarbons in food.* SCF/CS/CNTM/PAH/29 Final. Brussels, Belgium: European Commission. https://ec.europa.eu/food/sites/food/files/safety/docs/sci-com_scf_out153_en.pdf [accessed 24 May 2021].

EFSA (European Food Safety Authority). 2008. Polycyclic Aromatic Hydrocarbons in Food. Scientific Opinion of the Panel on Contaminants in the Food Chain (Question No. EFSA-Q-2007-136). EFSA J 724:1–114, https://doi.org/10.2903/j.efsa.2008.724.

EU (European Union). 2019. Directive (EU) 2019/130 of the European Parliament and of the Council of 16 January 2019 amending Directive 2004/37/EC on the protection of workers from the risks related to exposure to carcinogens or mutagens at work. Off J Eur Union L 30:112–120, http://data.europa.eu/eli/dir/2019/130/oj.

Goodman J, Lynch H. 2017. Improving the International Agency for Research on Cancer's consideration of mechanistic evidence. Toxicol Appl Pharm 319:39–46, PMID: 28162991, https://doi.org/10.1016/j.taap.2017.01.020.

Goodson WH III, Lowe L, Carpenter DO, Gilbertson M, Manaf Ali A, Lopez de Cerain Salsamendi A, et al. 2015. Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: the challenge ahead. Carcinogenesis 36(suppl 1):S254–S296, PMID: 26106142, https://doi.org/10.1093/carcin/bgv039.

Guo Y, Séaghdha DO, Silins I, Sun L, Högberg J, Stenius U, et al. 2014. CRAB2.0: a text mining tool for supporting literature review in chemical cancer risk assessment. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations.* Stroudsburg, PA: Association for Computational Linguistics, 76–80.

Gwinn MR, Axelrad DA, Bahadori T, Bussard D, Cascio WE, Deener K, et al. 2017. Chemical risk assessment: traditional vs. public health perspectives. Am J Public Health 107(7):1032–1039, PMID: 28520487, https://doi.org/10.2105/AJPH.2017.303771.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. Cell 100(1):57–70, PMID: 10647931, https://doi.org/10.1016/s0092-8674(00)81683-9.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. Cell 144(5):646–674, PMID: 21376230, https://doi.org/10.1016/j.cell.2011.02.013.

Harmston N, Filsell W, Stumpf MPH. 2010. What the papers say: text mining for genomics and systems biology. Hum Genomics 5(1):17–29, PMID: 21106487, https://doi.org/10.1186/1479-7364-5-1-17.

Hattis D, Chu M, Rahmioglu N, Goble R, Verma P, Hartman K, et al. 2009. A preliminary operational classification system for nonmutagenic modes of action for carcinogenesis. Crit Rev Toxicol 39(2):97–138, PMID: 19009457, https://doi.org/10.1080/10408440802307467.

IARC (International Agency for Research on Cancer). 2016. List of classifications. Agents classified by the IARC Monographs, Volumes 1–115. https://monographs.iarc.who.int/list-of-classifications [accessed 24 May 2021].

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. 2010. Some non-heterocyclic polycyclic aromatic hydrocarbons and some related exposures. IARC Monogr Eval Carcinog Risks Hum 92:1–853, PMID: 21141735.

Jarvis IWH, Dreij K, Mattsson Å, Jernström B, Stenius U. 2014. Interactions between polycyclic aromatic hydrocarbons in complex mixtures and implications for cancer risk assessment. Toxicology 321:27–39, PMID: 24713297, https://doi.org/10.1016/j.tox.2014.03.012.

Khalek IA, Blanks MG, Merritt PM, Zielinska B. 2015. Regulated and unregulated emissions from modern 2010 emissions-compliant heavy-duty on-highway diesel engines. J Air Waste Manag Assoc 65(8):987–1001, PMID: 26037832, https://doi.org/10.1080/10962247.2015.1051606.

Korhonen A, Séaghdha DO, Silins I, Sun L, Högberg J, Stenius U. 2012. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PLoS One 7(4):e33427, PMID: 22511921, https://doi.org/10.1371/journal.pone.0033427.

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, PMID: 19772619, https://doi.org/10.1186/1471-2105-10-303.

Krewski D, Bird M, Al-Zoughool M, Birkett N, Billard M, Milton B, et al. 2019. Key characteristics of 86 agents known to cause cancer in humans. J Toxicol Environ Health B Crit Rev 22(7–8):244–263, PMID: 31637961, https://doi.org/10.1080/10937404.2019.1643536.

Leung CC, Yu ITS, Chen W. 2012. Silicosis. Lancet 379(9830):2008–2018, PMID: 22534002, https://doi.org/10.1016/S0140-6736(12)60235-9.

Li R, Zhao L, Zhang L, Chen M, Dong C, Cai Z. 2017. DNA damage and repair, oxidative stress and metabolism biomarker responses in lungs of rats exposed to ambient atmospheric 1-nitropyrene. Environ Toxicol Pharmacol 54:14–20, PMID: 28668703, https://doi.org/10.1016/j.etap.2017.06.009.

Luch A. 2009. On the impact of the molecule structure in chemical carcinogenesis. EXS 99:151–179, PMID: 19157061, https://doi.org/10.1007/978-3-7643-8336-7_6.

Misaki K, Takamura-Enya T, Ogawa H, Takamori K, Yanagida M. 2016. Tumour-promoting activity of polycyclic aromatic hydrocarbons and their oxygenated or nitrated derivatives. Mutagenesis 31(2):205–213, PMID: 26656082, https://doi.org/10.1093/mutage/gev076.

NTP (National Toxicology Program). 2012. Research concept: Polycyclic aromatic hydrocarbons (PAHs)-draft. NTP Board of Scientific Counselors meeting, 11 December 2012. Research Triangle Park, NC: NTP, National Institute of Environmental Sciences.

NTP. 2016. *14th Report on Carcinogens*. Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service. https://ntp.Niehs.Nih.Gov/go/roc14 [accessed 24 May 2021].

Papamokos G, Silins I. 2016. Combining QSAR modeling and text-mining techniques to link chemical structures and carcinogenic modes of action. Front Pharmacol 7:284, PMID: 27625608, https://doi.org/10.3389/fphar.2016.00284.

Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. 2012. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet 13(12):829–839, PMID: 23150036, https://doi.org/10.1038/nrg3337.

Sakai A, Sasaki K, Muramatsu D, Arai S, Endou N, Kuroda S, et al. 2010. A Bhas 42 cell transformation assay on 98 chemicals: the characteristics and performance for the prediction of chemical carcinogenicity. Mutat Res 702(1):100–122, PMID: 20656056, https://doi.org/10.1016/j.mrgentox.2010.07.007.

Satpathy SR, Jala VR, Bodduluri SR, Krishnan E, Hegde B, Hoyle GW, et al. 2015. Crystalline silica-induced leukotriene B4-dependent inflammation promotes lung tumour growth. Nat Commun 6:7064, PMID: 25923988, https://doi.org/10.1038/ncomms8064.

Silins I, Korhonen A, Högberg J, Stenius U. 2012. Data and literature gathering in chemical cancer risk assessment. Integr Environ Assess Manag 8(3):412–417, PMID: 22275076, https://doi.org/10.1002/ieam.1278.

Silins I, Korhonen A, Stenius U. 2014. Evaluation of carcinogenic modes of action for pesticides in fruit on the Swedish market using a text-mining tool. Front Pharmacol 5:145, PMID: 25002848, https://doi.org/10.3389/fphar.2014.00145.

Smith MT, Guyton KZ, Gibbons CF, Fritz JM, Portier CJ, Rusyn I, et al. 2016. Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. Environ Health Perspect 124(6):713–721, PMID: 26600562, https://doi.org/10.1289/ehp.1509912.

Solano MLM, Rocha NS, Barbisan LF, Franchi CAS, Spinardi-Barbisan ALT, de Oliveira MLCS, et al. 2016. Alternative multiorgan initiation–promotion assay for chemical carcinogenesis in the Wistar rat. Toxicol Pathol 44(8):1146–1159, PMID: 28245158, https://doi.org/10.1177/0192623316678931.

U.S. EPA (U.S. Environmental Protection Agency). 1993. *Provisional Guidance for Quantitative Risk Assessment of Polycyclic Aromatic Hydrocarbons*. EPA/600/R-93/089. Washington, DC: U.S. EPA. https://semspub.epa.gov/work/HQ/100000047.pdf [accessed 25 May 2021].

U.S. EPA. 2010a. *Development of Relative Potency Factor (RPF) Approach for Polycyclic Aromatic Hydrocarbon (PAH) Mixtures*. External review draft. EPA/635/R-08/012A. Washington, DC: U.S. EPA.

U.S. EPA. 2010b. *Development of a Relative Potency Factor (RPF) Approach for Polycyclic Aromatic Hydrocarbon (PAH) Mixtures: In Support of Summary Information on the Integrated Risk Information System (IRIS)*. External review draft. EPA/635/R-08/012A. Washington, DC: U.S. EPA. https://yosemite.epa.gov/sab/sabproduct.nsf/0/E65D909C98520C1D85257501005E46AE/$File/IRIS_PAH_RPF_ERD_Feb+2010.pdf [accessed 25 May 2021].

Vermeulen R, Rothman N, Lan Q. 2011. Coal combustion and lung cancer risk in XuanWei: a possible role of silica? Med Lav 102(4):362–367, PMID: 21834273.

WHO (World Health Organization). 2000. *Air Quality Guidelines for Europe*. 2nd ed. European Series, No. 91. Copenhagen, Denmark: WHO Regional Publications.

Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. 2013. Biomedical text mining and its applications in cancer research. J Biomed Inform 46(2):200–211, PMID: 23159498, https://doi.org/10.1016/j.jbi.2012.10.007.