Natural language processing in toxicology: Delineating adverse outcome pathways and guiding the application of new approach methodologies

Marie P.F. Corradi [a,*], Alyanne M. de Haan [a], Bernard Staumont [b], Aldert H. Piersma [c], Liesbet Geris [b], Raymond H.H. Pieters [a], Cyrille A.M. Krul [a], Marc A.T. Teunis [a,*]

[a] *Innovative Testing in Life Sciences and Chemistry, University of Applied Sciences Utrecht, Heidelberglaan 7, Utrecht 3584 CS, the Netherlands*
[b] *Biomechanics Research Unit, GIGA In Silico Medicine, University of Liège, Avenue de l'Hôpital 11, Liège 4000, Belgium*
[c] *Centre for Health Protection of the Dutch National Institute for Public Health and the Environment (RIVM), Heidelberglaan 8, Utrecht 3584 CS, the Netherlands*

A B S T R A C T

Adverse Outcome Pathways (AOPs) are conceptual frameworks that tie an initial perturbation (molecular initiating event) to a phenotypic toxicological manifestation (adverse outcome), through a series of steps (key events). They provide therefore a standardized way to map and organize toxicological mechanistic information. As such, AOPs inform on key events underlying toxicity, thus supporting the development of New Approach Methodologies (NAMs), which aim to reduce the use of animal testing for toxicology purposes.

However, the establishment of a novel AOP relies on the gathering of multiple streams of evidence and information, from available literature to knowledge databases. Often, this information is in the form of free text, also called unstructured text, which is not immediately digestible by a computer. This information is thus both tedious and increasingly time-consuming to process manually with the growing volume of data available. The advancement of machine learning provides alternative solutions to this challenge. To extract and organize information from relevant sources, it seems valuable to employ deep learning Natural Language Processing techniques.

We review here some of the recent progress in the NLP field, and show how these techniques have already demonstrated value in the biomedical and toxicology areas. We also propose an approach to efficiently and reliably extract and combine relevant toxicological information from text. This data can be used to map underlying mechanisms that lead to toxicological effects and start building quantitative models, in particular AOPs, ultimately allowing animal-free human-based hazard and risk assessment.

1. Introduction

Guaranteeing the health of humans means that every product coming on the market needs to be thoroughly tested for possible adverse effects on human health. In the past, this has largely been done with the help of animal testing. However, animal testing raises ethical issues and is not a sustainable evaluation method considering the time and financial investments as well as the growing number of substances to evaluate, be it for agricultural, cosmetic, food, medical or energy applications. For that purpose, a new class of evaluation methods, the New Approach Methodologies (NAMs), have emerged. NAMs refer to any non-animal-based method, whether *in vitro* or *in silico* [1].

Adverse Outcome Pathways (AOPs) are a standardized description of toxicological mechanistic information and can form the basis for development of New Approach Methodologies (NAMs). These mechanistic frameworks link molecular triggers to an adverse outcome via a series

of steps covering various biological levels, and depend on a wealth of data to provide strong evidence of the relationships between each step [2]. Traditionally, this relies heavily on animal testing. However, several ongoing projects focus on making better use of the knowledge already available in literature, in particular ONTOX [43] and VHP4Safety [45]. Information extraction was previously mostly done manually, while the body of evidence in scientific literature is growing every year. Therefore, new methods are being developed to scan existing literature and extract relevant knowledge from it using machine learning. This paper is focusing on the potential of automated text analysis, better known as Natural Language Processing (NLP), to support and facilitate the process of knowledge extraction for AOP building, leading to the development of NAMs.

The recent advances in NLP and the way they can be applied to support AOPs development will be discussed in the scope of current projects ONTOX and VHP4Safety.

2. Adverse outcome pathways in modern toxicology

2.1. What are AOPs?

An AOP is a framework to organize existing knowledge pertaining to the mechanisms by which adverse effects are triggered in human organisms [2]. It spans various levels of biological organization, starting with a Molecular Initiating Event (MIE) and ending with one or multiple Adverse Outcome(s) (AO) at the organ, organism or population level. An outcome is considered as AO if it is relevant for hazard and risk assessment. The MIE and AO are connected by a series of molecular, cellular or organ-level events called Key Events (KE). Each pair of KE is connected by a Key Event Relationship (KER) that describes the relationship between the upstream and downstream event and can be causal, mechanistic, inferential or correlation-based. The KER therefore supports and gives quantitative insight for the extrapolation of the downstream KE from the state of the upstream KE [3]. In order to prove the relation between KEs, KERs rely on a wealth of information to feed the weight of evidence (WoE) supporting their existence. In that scope, AOPs can be used to identify critical missing information for hazard and risk assessment. AOP developers can also identify endpoints that can be assessed *in vitro* or *in silico* by a NAM to estimate risk for human, hence contributing to a transition towards animal-free testing [4].

Of course, in reality toxicological mechanisms are more complex. They are neither linear nor one-directional, but rather a network of interactions at the molecular, cellular and organ level, including feedback loops that allow to understand why lower doses of a compound are not toxic despite triggering a MIE. AOPs are therefore intended to be modular and combinable in AOP networks, as in Spinu et al. [5].

AOPs are applicable to multiple species, by focusing on mechanistic relations between KEs rather than apical endpoints in specific species. Indeed, identifying conserved molecular events between species will help regulators extrapolate effects thus reducing the number of animals needed to predict adverse outcomes in humans [6].

Moreover, AOPs are stressor-agnostic: the first element of an AOP is a MIE. This event can be triggered by one or multiple chemicals or other stressors (e.g., radiation, inflammation), but is not defined by a single one, making AOPs much more flexible than previous approaches [3].

2.2. AOPs and NAMs

On the chemical level, structure-based models are used to predict chemical activity, such as QSARs (Quantitative Structure-Activity Relationship) [7], but also to evaluate chemical similarity and allow grouping of chemicals, such as RA (Read-Across) [8]. RASAR (Read-Across Structure-Activity Relationship), a combination of the previous two approaches, has also showed promising results. RASAR uses chemical hazard properties as well as structural properties to predict hazard [9]. These methods can be combined with AOPs to allow for more efficient screening of compounds by determining which chemicals are likely to trigger the MIE based on their structural properties. They can therefore help rule entire groups of chemicals in or out for testing, without the need for additional data. By drastically decreasing the amount of chemicals to test, these approaches support the reduction of required animal testing and therefore fall under the denomination of NAMs.

2.3. Current approaches building AOPs

The construction of an AOP relies heavily on the gathering and reviewing of extensive data and literature, in order to support confidence and adoption [10]. The OECD [11] has suggested the adoption of tailored Bradford-Hill (BH) considerations to assess the confidence in an AOP. The BH criteria were originally designed to support evaluation of causality of relationships for epidemiological evidence and were adapted to fit the AOP framework. In Becker et al. [12], these criteria are referred to as:

1 Biological plausibility of KERs: what is the level of understanding and acceptance of the biology behind a given KER?
2 Empirical support for KERs: dose-response, temporality, and incidence
3 Essentiality of (KEs): blocking an upstream KE results in the downstream KE not occurring. This can also be proved reversibly, in other words removing the block on the upstream KE leads to the downstream KE happening again.

These criteria are demonstrated in the article by Horvat et al. [13], describing the AOP leading to liver fibrosis from protein alkylation. Horvat et al. [13] followed a top-down approach, selecting an adverse outcome of interest and tracing it back to the key events and MIE(s) of interest. Though AOPs are stressor-agnostic, the authors have used chemicals clinically known to trigger the AO of interest, namely carbon tetrachloride and allyl alcohol, in order to understand the mechanistic process behind the biological response. An overview of all the evidence supporting each KE or KER is also provided, ranging between two and nine supporting scientific articles for each item, totaling close to 180 references for this paper. The authors most likely first had to screen existing literature to select relevant articles. A PubMed search using the keywords liver fibrosis AND (carbon tetrachloride OR allyl alcohol), and limiting the results to before 2017, yielded more than 3500 results https://pubmed.ncbi.nlm.nih.gov/[44]. Reviewing so many papers manually was evidently very cumbersome.

It is clear that AOP building depends on the availability and organization of data. The amount and quality of data used is actually critical to drive confidence and endorsement from peers as well as support usage [14]. Because AOPs are based on manually constructed relationships, WoE determinations are pivotal for their acceptance and usefulness. The BH criteria adapted for application in the AOP framework are a good starting point when considering (semi-)automated construction of AOPs and any computational approach should be built in concordance. Kleinstreuer et al. [15] have already suggested the potential of data science to help provide suggestions of AOP associations from scientific literature [15]. Here we propose to use Natural Language Processing (NLP) to extract AOP-relevant relationships, in concordance with the BH criteria, as one way of moving beyond manual construction of AOPs and into computer-aided extraction of existing knowledge from text.

3. Advances in Natural Language Processing

3.1. Natural Language Processing techniques and evolution

Machine learning refers to the ability of a computer to learn from data, rather than being explicitly programmed to accomplish a given task [16]. This is often used to mimic human behavior: for example, recognizing a specific person on a picture, understanding a piece of text or determining a causal relationship between two events [17]. Harvesting computing power can multiply scientific discoveries by analyzing large amounts of data and finding previously unknown relationships or knowledge. Machine learning starts with data as an input, and learns a mathematical model on this data. As it operates on mathematical transformations, any input to a machine learning model is numerical.

NLP is a machine learning field concerned with automatically analyzing human language. It has broad application, from characterizing sentiments and detecting hate speech in social media [18,19] to automatic machine translation from one human language to another [20]. When building AOPs, NLP could be used to programmatically retrieve scientific articles, fragment them into paragraphs of interest and automatically mine the relevant relationships between MIEs, KEs and AOs. This information then needs to be stored in a machine-readable way. The AOP developer's role could then focus more on quality control and critical evaluation of the information extracted. In this review paper, we are particularly interested in the progress of these NLP techniques

	Administration	of	valproic	acid	during	first	trimester	of	pregnancy	causes	neural	tube	defects
Tokenization	Administration	of	valproic	acid	during	first	trimester	of	pregnancy	causes	neural	tube	defects
PoS	noun	adp	adj	noun	adp	adj	noun	adp	noun	verb	adj	noun	noun
NER	–		chemical								adverse outcome		
Entity Linking			PubChem CID: 3121								HP:0045005 *Neural tube defect*		
Dependency parsing/ RelEx			cause								effect		

Fig. 1. Common NLP tasks, applied to a biological example. PoS: Part-of-Speech tagging, NER: Named Entity Recognition, RelEx: Relation extraction, adj: adjective, adp: adposition, CID: chemical ID, HP: Human Phenotype Ontology term.

pertinent to extracting information from written text, and specifically the progress of neural networks.

NLP can be broken down into several tasks following each other. First, the text of interest is broken into words according to a set of rules depending on the language (grammar, punctuation, etc.) as well as the topic discussed (abbreviations). This first step is called tokenization [21]. Secondly, Part-of-Speech (PoS) tagging maps each word of a sentence to its grammatical function, e.g., noun, verb, adjective, etc. From there, dependency parsing allows to determine the semantic relationships between words, which can potentially lead to discovering relationships between them. So-called "higher level" tasks also build on these preliminary steps [22]. We show an example of some of these tasks applied to a sentence in a biological context in Fig. 1. We see in particular three tasks of interest for information extraction for AOP building:

1 Named Entity Recognition (NER) to recognize concepts such as MIE, KEor AO.
2 Relation Extraction to support KERs: causal relationships feed the WoE.
3 Entity linking assigns a unique identifier to a given (group of) word(s). The advantage is two-fold: disambiguating the word as well as allowing further relations to e.g. existing databases. In Fig. 1, valproic acid is for example linked to its PubChem chemical identifier (CID) [23]. This identifier not only allows the reader to associate the term "valproic acid" with its structure but also to identify synonyms (e.g., "valproate" here) with the same PubChem CID. The information obtained on valproate and valproic acid is then linked as pertaining to the same chemical.

Most NLP tasks can be described as a classification task: assigning a label x to an input text. For example, NER maps a (group of) word(s) ("valproic acid") to a concept ("chemical") (Fig. 1). Language processing can further be viewed as a sequence modeling problem: in order to understand a specific part or word of the sentence, we need to be aware of what comes before and after it. For example, the word "expression" can refer among others to a mathematical formula (the expression of the area of a rectangle), the process of translating the information contained in a gene (gene expression) or the emotion conveyed on one's face (facial expression). However, in isolation, it is almost impossible to distinguish which of these meanings the author wants to convey: the context is therefore key to disambiguate. In that regard, some significant progress came from the application of neural networks to language modeling,

A neural network (sometimes also called deep learning) is a type of machine learning architecture built by analogy to the biological concept of a network of neurons. In a neural network, a number of neurons, also called nodes, are densely connected in layers. Each neuron receives a signal from its environment. This environment can be other neurons or input data, and different sources from the environment can have different weights. All the input a given neuron receives is combined via a mathematical transformation. If the combined input is higher than the activation threshold of the neuron, an output is transmitted to the next layer of neurons [24]. The combinatorial possibilities are diverse in terms of number of neurons, number of layers, connection between neurons, etc., and define different architectures [25]. These various architectures can learn complex, non-linear functions mapping an input data (e.g., a sequence of words) to an output of interest (e.g., identification of "expression of gene X" as a KE).

As mentioned previously, the input to any machine learning model is numerical. This means that for text analysis, the first step is to translate words into numbers. One of the major innovations for NLP has been the mapping of words to vectors representing their semantic and contextual similarity [26]. In other terms, words representing concepts that have a similar meaning or are often used together will have similar encodings. These word vectors are usually learned using deep learning on corpora comprising millions of texts such as Common Crawl, a large corpus of several millions of news articles [27].

Word vectors can then be fed into neural networks architecture optimized for sequence modeling, in particular Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM). CNNs, traditionally used in computer vision, have a particular element of architecture named the convolutional layer. This layer allows to extract features gathering input at different positions of a vector. In other words, this allows to consider "sliding windows" of words across a sentence [28]. On the other hand, LSTMs incorporate a layer allowing to access previous states, in other terms, to remember words coming previously in a sentence. In addition, they include a possibility to forget part of what has already been seen [29]. These advances in neural network architectures improved language models by capturing longer term dependencies and improving sequence modeling of text.

However, these models are still built as a pipeline of models trained and applied one after the other. There is value in sharing information between one task and the next. Information used for PoS tagging, for example, can be equally useful for dependency parsing as they are both relying on grammatical properties. More recently, a new architecture for

language models has emerged: the Transformer [30]. Transformer models comprise an attention mechanism, which accounts for the relationships and dependencies between all the words in a sentence, no matter their distance to each other. It assigns differential weights to all words in relation with individual words, allowing to indicate which words are essential for the interpretation of others by giving them a higher weight. The difference with the word vectors concept we presented above is that word vectors are trained on a general corpus. Here, weights are different in each sentence according to the context. This means that the context of a word can be re-used for each task, instead of being learnt multiple times. Transformers have shown very good performance on translation tasks in this founding paper, so it is very likely they will outperform traditional architectures on other NLP tasks as well.

A lot of progress has been made in the NLP field. We will now review how much of this progress has been applied to the field of toxicology.

3.2. Current NLP use in the biomedical and toxicology fields

There have been a number of (collective) efforts on the development of language models and applications for the biomedical field in the last ten years [31,32]. Pletscher-Frankild et al. [33] have developed a system to extract gene-disease associations from literature, based on dictionary mapping and co-occurrence. LimTox [34] is a full pipeline and web application designed to retrieve scientific text and extract associations between compounds and toxicological endpoints, focused on hepatotoxicity. However, it is very specifically focused on the extraction of relations between biochemical markers and cytochromes and does not extend to other molecular interactions nor to relations between molecular events and higher level events such as organ or organism-level events. It is also very much rule and pattern-based and does not seem to take full advantage of the promises of NLP for scientific text.

More recently, Minet et al. [35] built ad-hoc gene sets from the mucus hypersecretion AOP developed by Luettich et al. [36]. Publications related to each KE from this AOP were gathered from PubMed, and a list of "seed" genes per KE were extracted manually from this primary corpus. These formed the basis for an extended search which associated these genes with AOP-associated terms such as "smoke" or "tobacco." Further genes were extracted by NER, and co-occurrence between each pair of genes across the corpus was calculated. After manual curation, the gene set for each KE could be used in gene expression assays to discriminate e.g. samples exposed to cigarette smoke. This work is an example of how NLP can also support the development of new *in vitro* tests to evaluate toxicologically relevant endpoints as well as identify gaps in data.

Jornod et al. [37] have taken NLP one step closer to AOP development by building AOP-helpFinder, a web platform to find existing literature connecting stressors and biological events. This platform is based on the input of existing AOP elements such as MIE/KE/AO. The platform then searches for these elements in the text of existing publications. This search is based on the matching of (a simplified version of the) words comprising the MIE/KE/AO of interest in the text, and calculating a score based on the shortest path between these words. In other words, the closer these terms are in the text, the more likely it is that the tool has correctly identified an event [38]. To the best of our knowledge, this tool identifies events and determines whether they are related by looking if they are co-occurring in text, but is not making use of semantic information to determine causal relationships between events. In addition, this approach seems to mostly be valuable for the extraction of knowledge about already established events, as the user needs to input MIE/KE/AO from AOPWiki or other sources. Therefore, while the tool has established value to uncover new evidence for KERs between existing events [39], the discovery of new events, in particular MIEs or KEs, seems somewhat more restricted.

All these are interesting applications but do not necessarily take advantage of the recent technical progress in NLP. These recent advances have been mostly implemented in the field of pharmacovigilance. Weis-senbacher et al. [40] have used deep neural networks, notably LSTMs, to find mentions of medications in tweets. This could be used further to detect those associated with mentions of adverse effects or toxicities. Wang et al. [41] have used a Transformer-based model on case reports from the FDA Adverse Event Reporting System to predict causal relationships between analgesics treatment and liver failure as well as between Tramadol intake and mortality. They report a better performance, and in particular identify more causal factors, than the traditional signal detection methods. This is very valuable, and is getting closer to harvesting the power of NLP for NAMs. However, it is still targeting a restricted part of toxicology by directly linking compounds and Adverse Outcomes without exploring intermediate mechanistic steps.

From these examples, it is clear that some NLP tasks and architectures have been applied in the biomedical and in particular in the toxicology domain for knowledge extraction, and some more advanced techniques for related domains. We think the recent technical progress in NLP can be used further for several aspects of AOPs and NAMs development, and propose our approach in the following section.

4. Future of Natural Language Processing in the toxicology field

4.1. Opportunities of NLP for toxicology

We have seen the value that NLP can bring to extract information for toxicological purposes. In Fig. 2, we show the different ways NLP can help NAMs development, with a focus on the AOP framework.

First, connecting chemicals to molecular, cellular, organ or organism-level events can enrich approaches such as RASAR with biological information in addition to the purely physico-chemical parameters already used. The automatic acquisition and addition of that information will then support a more accurate and fine-grained chemical selection. In addition, it can direct toxicological assessment to AOP network areas expected to be triggered by the compound(s), or even to prioritize compounds to be tested, by selecting compounds presenting existing evidence of triggering an AO of interest to test first.

Second, we envision NLP as a support tool for BH criteria. We suggest it can help gather weight of evidence about KERs and support their biological plausibility. For example, highlighting causal relationships between upstream (molecular) key events and downstream (cellular) key events will considerably facilitate the review process by AOP developers. The automation of literature search will also highlight data gaps more efficiently than manual review, which could be filled in by the development of new testing methods (NAMs) as evidenced earlier [35].

Finally, we see NLP as an additional tool to evaluate conservation between species. By extracting biological events and linking them to formally organized concepts such as e.g. ontology terms from the Mammalian Phenotype Ontology [42], we can derive what is conserved in humans from other species for which data exists.

4.2. Enacting NLP for toxicology

We described above how we believe NLP can support the toxicology field. Practically, we suggest a workflow for information extraction, shown in Fig. 3.

The first step is a programmatic query of the relevant literature with well-defined keywords. This can either be a search related to toxic effects of selected chemicals (top-down) or to an AO of interest (bottom-up). As this automatic search will produce an excessive number of results, a curation step is needed. This can be supported by NLP, for example by using NER to recognize biological entities of interest in the text. From then on, the text from selected papers can be processed with the steps we described earlier in Fig. 1. This automatic extraction of information from potentially hundreds or thousands of articles will in our opinion contribute to a more efficient gathering of evidence than manual reviews. This step can and should also comprise some manual
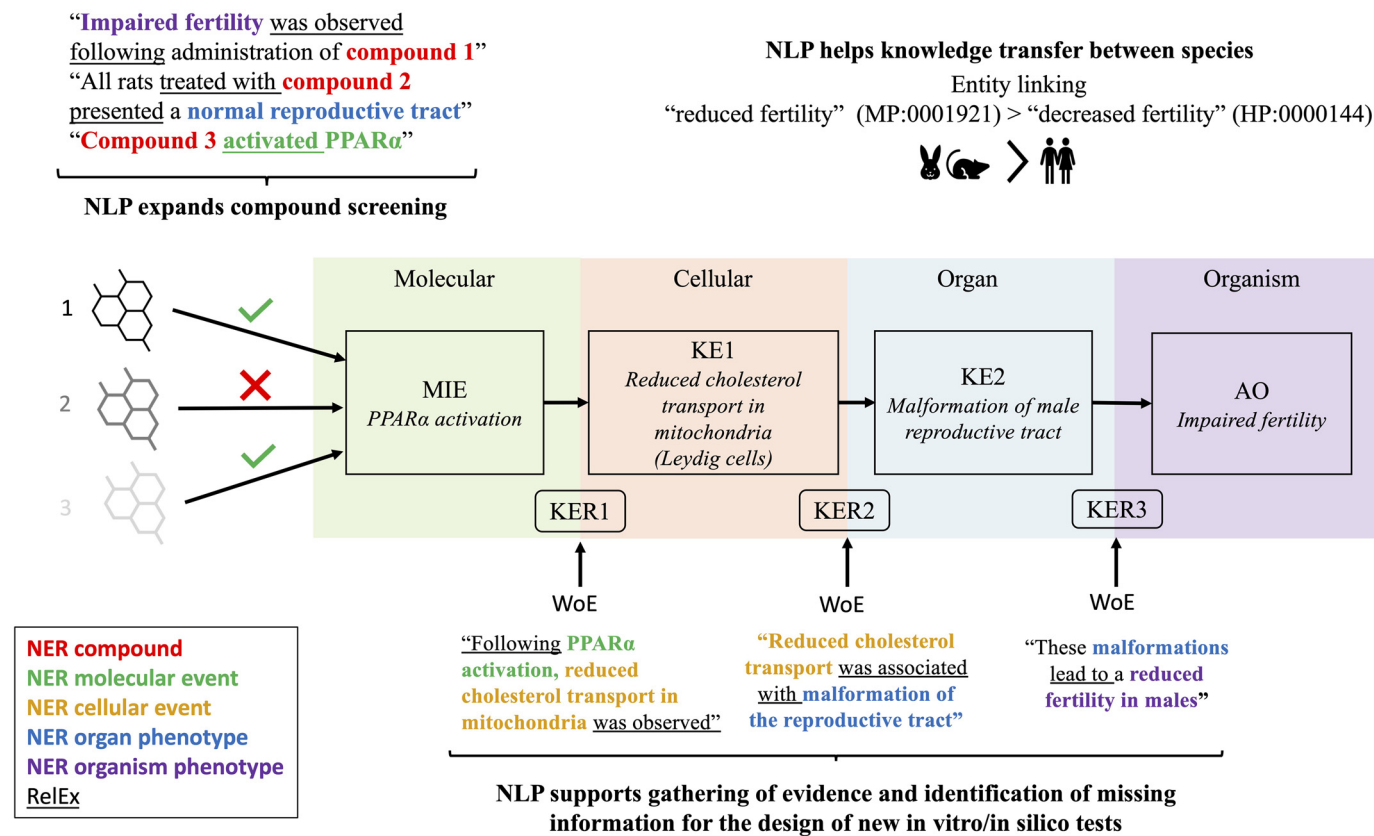
Fig. 2. NLP support for AOPs. Here we give an example on how NLP can support the development of AOP with the example of AOP 18 in AOPWiki, PPARalpha activation in utero leading to impaired fertility in males [46]. The example sentences we are giving are meant for illustration and not extracted from any specific article. It should also be noted that we provide one example for each KER, but a lot more would be needed to strengthen and give confidence in the WoE. In addition, by gathering evidence on mechanistic relations between key events that can be present in multiple species, NLP can facilitate the transfer of knowledge between animal models, as well as predict its relevance for human health. MIE: Molecular Initiating Event, KE: Key Event, KER: Key Event Relationship, AO: Adverse Outcome, WoE: Weight of Evidence, NER: Named Entity Recognition, RelEx: Relation Extraction.
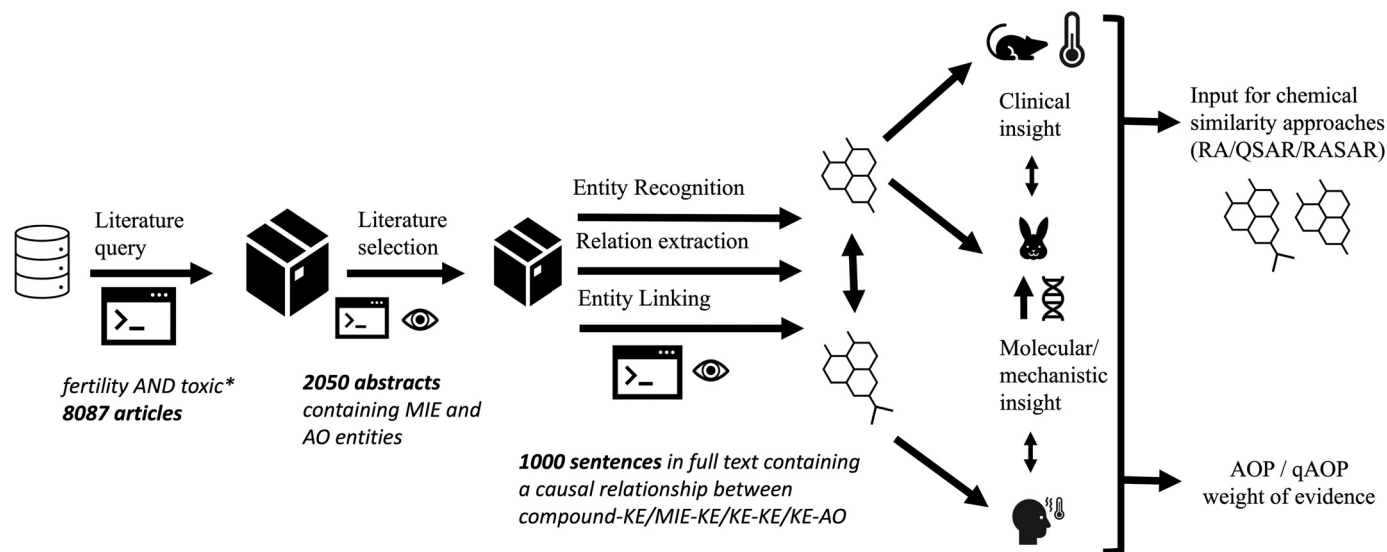
Fig. 3. NLP workflow to support gathering of toxicological insights. We give an example of Pubmed query and the number of articles yielded, and how this could be reduced to relevant information (numbers given in the second and third steps of the workflow are theoretical and meant to show how the filtering can be helped by NLP). Example sentences can be found in Fig. 2. This data could then be used for multiple use cases: evaluation of conservation of mechanisms between species, weight of evidence for the establishment of (quantitative) AOPs, or data enrichment for chemical selection process.

inspection from the AOP developer to ensure the models are providing correct information.

All the information is gathered quickly and can be stored immediately in machine-readable format. We envision storing information into graph databases, which would allow easy visualization of relationships and their WoE, for example to support biological plausibility of KERs as exemplified in Fig. 2. It can also be further processed to support NAMs as described in Fig. 2. For example, it can be used as input for similarity approaches to help in chemical screening: chemicals that trigger similar events according to literature could be grouped together.

We believe that this workflow is flexible enough to help the discovery of novel events as well as feed the WoE for KERs within existing AOPs. It can also facilitate combining AOPs into AOP networks.

We see this approach as possibly limited by two factors: the availability of publications in open-science format, as well as the bias for positive results in publications. However, these biases would be similar with a non-automated scheme. We strongly believe that this workflow will be more efficient than a manual approach by accelerating review as well as allowing the evaluation of gaps in currently available knowledge.

5. Conclusion

We show here that animal testing-based toxicology can gain from the introduction of new machine learning techniques, in particular NLP, to make better use of the information available in scientific literature. NLP can help describe AOPs, and thereby support the selection of KEs that should be tested in NAMs, driving the selection and novel development of NAMs and paving the way for human-based animal-free chemical hazard and risk assessment.

The workflow we presented here can also be applied to other areas of biomedicine and even Life Sciences in general. We propose that whenever information extraction from an existing body of literature is required, NLP could be applied in this fashion to improve the knowledge gathering process.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

References

[1] Punt A, Bouwmeester H, Blaauboer BJ, Coecke S, Hakkert B, Hendriks DFG, Jennings P, Kramer NI, Neuhoff S, Masereeuw R, Paini A, Peijnenburg AACM, Rooseboom M, Shuler ML, Sorrell I, Spee B, Strikwold M, Van der Meer AD, Van der Zande M, Vinken M, Yang H, Bos PMJ, Heringa MB. New approach methodologies (NAMs) for human-relevant biokinetics predictions. Meeting the paradigm shift in toxicology towards an animal-free chemical risk assessment. ALTEX 2020;37:607–22. doi:10.14573/altex.2003242.

[2] Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, Serrrano JA, Tietge JE, Villeneuve DL. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 2010;29:730–41. doi:10.1002/etc.34.

[3] Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B, Lettieri T, Munn S, Nepelska M, Ottinger MA, Vergauwen L, Whelan M. Adverse outcome pathway (AOP) development I: strategies and principles. Toxicol Sci 2014;142:312–20. doi:10.1093/toxsci/kfu199.

[4] Clippinger AJ, Allen D, Behrsing H, BéruBé KA, Bolger MB, Casey W, DeLorme M, Gaça M, Gehen SC, Glover K, Hayden P, Hinderliter P, Hotchkiss JA, Iskandar A, Keyser B, Luettich K, Ma-Hock L, Maione AG, Makena P, Melbourne J, Milchak L, Ng SP, Paini A, Page K, Patlewicz G, Prieto P, Raabe H, Reinke EN, Roper C, Rose J, Sharma M, Spoo W, Thorne PS, Wilson DM, Jarabek AM. Pathway-based predictive approaches for non-animal assessment of acute inhalation toxicity. Toxicol Vitr 2018;52:131–45. doi:10.1016/j.tiv.2018.06.009.

[5] Spinu N, Bal-Price A, Cronin MTD, Enoch SJ, Madden JC, Worth AP. Development and analysis of an adverse outcome pathway network for human neurotoxicity. Arch Toxicol 2019;93:2759–72. doi:10.1007/s00204-019-02551-1.

[6] Burden N, Sewell F, Andersen ME, Boobis A, Chipman JK, Cronin MTD, Hutchinson TH, Kimber I, Whelan M. Adverse Outcome Pathways can drive non-animal approaches for safety assessment. J Appl Toxicol JAT 2015;35:971–5. doi:10.1002/jat.3165.

[7] Khan PM, Roy K. Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR). Exp Opin Drug Discov 2018;13:1075–89. doi:10.1080/17460441.2018.1542428.

[8] Patlewicz G, Ball N, Boogaard PJ, Becker RA, Hubesch B. Building scientific confidence in the development and evaluation of read-across. Regulatory toxicology and pharmacology. RTP 2015;72:117–33. doi:10.1016/j.yrtph.2015.03.015.

[9] Luechtefeld T, Marsh D, Rowlands C, Hartung T. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol Sci 2018;165:198–212. doi:10.1093/toxsci/kfy152.

[10] Perkins EJ, Ashauer R, Burgoon L, Conolly R, Landesmann B, Mackay C, Murphy CA, Pollesch N, Wheeler JR, Zupanic A, Scholz S. Building and applying quantitative adverse outcome pathway mod-els for chemical hazard and risk assessment. Environ Toxicol Chem 2019;38:1850–65. doi:10.1002/etc.4505.

[11] OECD Users' handbook supplement to the guidance document for developing and Assessing adverse outcome pathways. OECD; 2018.

[12] Becker RA, Ankley GT, Edwards SW, Kennedy SW, Linkov I, Meek B, Sachana M, Segner H, Van Der Burg B, Villeneuve DL, Watanabe H, Barton-Maclaren TS. Increasing scientific confidence in adverse outcome pathways: application of tailored bradford-hill considerations for evaluating weight of evidence. Regul Toxicol Pharma RTP 2015;72:514–37. doi:10.1016/j.yrtph.2015.04.004.

[13] Horvat T, Landesmann B, Lostia A, Vinken M, Munn S, Whelan M. Adverse outcome pathway development from protein alkylation to liver fibrosis. Arch Toxicol 2017;91:1523–43. doi:10.1007/s00204-016-1814-8.

[14] Svingen T, Villeneuve DL, Knapen D, Panagiotou EM, Draskau MK, Damdimopoulou P, O'Brien JM. A pragmatic approach to adverse outcome pathway development and evaluation. Toxicol Sci 2021;184:183–90. doi:10.1093/toxsci/kfab113.

[15] Kleinstreuer NC, Sullivan K, Allen D, Edwards S, Mendrick DL, Embry M, Matheson J, Rowlands JC, Munn S, Maull E, Casey W. Adverse outcome pathways: from research to regulation scientific workshop report. Regul Toxicol Pharmacol RTP 2016;76:39–50. doi:10.1016/j.yrtph.2016.01.007.

[16] Cohen S. Chapter 2 - the basics of machine learning: strategies and techniques. In: Cohen S, editor. Artificial intelligence and deep learning in pathology. Elsevier; 2021. p. 13–40. doi:10.1016/B978-0-323-67538-3.00002-6.

[17] Shalev-Shwartz S, Ben-David S. Understanding machine learning: from theory to algorithms. Cambridge: Cambridge University Press; 2014. doi:101017/CBO9781107298019.

[18] Chong WY, Selvaretnam B, Soon LK. Natural language processing for sentiment analysis: an exploratory analysis on tweets. In: Proceedings of the 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology; 2014. p. 212–17. doi:10.1109/ICAIET.2014.43.

[19] Schmidt A, Wiegand M. A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Valencia, Spain: Association for Computational Linguistics; 2017. p. 1–10. doi:10.18653/v1/W171101.

[20] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:1609.08144 [cs]. doi:10.48550/arXiv.1609.08144

[21] Webster JJ, Kit C. Tokenization as the initial phase in NLP. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 4, COLING '92. USA: Association for Computational Linguistics; 1992. p. 1106–10. doi:10.3115/992424.992434.

[22] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18:544–51. doi:10.1136/amiajnl-2011-000464.

[23] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2020;49:D1388–95. doi:10.1093/nar/gkaa971.

[24] Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods Neural Comput Micrbiol 2000;43:3–31. doi:10.1016/S0167-7012(00)002013.

[25] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. doi:10.1038/nature14539.

[26] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs]. doi:10.48550/arXiv.1301.3781

[27] Mackenzie J, Benham R, Petri M, Trippas JR, Culpepper J, Moffat A. CC-news-en: a large English news corpus. CIKM 2020. doi:10.1145/3340531.3412762.

[28] Pham NQ, Kruszewski G, Boleda G. Convolutional neural network language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics; 2016. p. 1153–62. doi:10.18653/v1/D16-1123.

[29] Sundermeyer, M., Schlüter, R., Ney, H., Interspeech, 2012. LSTM neural networks for language modeling. doi:10.21437/Interspeech.2012-65

[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention is all you need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017.

[31] Kim JD, Wang Y, Takagi T, Yonezawa A. Overview of Genia event task in BioNLP shared task 2011. In: Proceedings of the BioNLP Shared Task 2011 Workshop. Portland, Oregon, USA: Association for Computational Linguistics; 2011. p. 7–15.

[32] Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task 319–327. Association for Computational Linguistics; 2019. doi:10.18653/v1/W19-5034.

[33] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease–gene associations. Methods Text Mining Biomed Lit 2015;74:83–9. doi:10.1016/j.ymeth.2014.11.020.

[34] Cañada A, Capella-Gutierrez S, Rabal O, Oyarzabal J, Valencia A, Krallinger M. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. Nucleic Acids Res 2017;45:W484–9. doi:10.1093/nar/gkx462.

[35] Minet E, Haswell LE, Corke S, Banerjee A, Baxter A, Verrastro I, De Abreu E Lima F, Jaunky T, Santopietro S, Breheny D, Gaça MD. Application of text mining to develop AOP-based mucus hypersecretion genesets and confirmation with *in vitro* and clinical samples. Sci Rep 2021;11:6091. doi:10.1038/s41598-021-85345-9.

[36] Luettich K, Talikka M, Lowe FJ, Haswell LE, Park J, Gaca MD, Hoeng J. The adverse outcome pathway for oxidative stress-mediated EGFR activation leading to decreased lung function. Appl *In Vitro* Toxicol 2017;3:99–109. doi:10.1089/aivt.2016.0032.

[37] Jornod F, Jaylet T, Blaha L, Sarigiannis D, Tamisier L, Audouze K. AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development. Bioinformatics 2022;38:1173–5. doi:10.1093/bioinformatics/btab750.

[38] Carvaillo JC, Barouki R, Coumoul X, Audouze K. Linking Bisphenol S to adverse outcome pathways using a combined text mining and systems biology approach. Environ. Health Perspect. 2019;127:47005. doi:10.1289/EHP4200.

[39] Jornod F, Rugard M, Tamisier L, Coumoul X, Andersen HR, Barouki R, Audouze K. AOP4EUpest: mapping of pesticides in adverse outcome pathways using a text mining tool. Bioinformatics 2020;36:4379–81. doi:10.1093/bioinformatics/btaa545.

[40] Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, GonzalezHernandez G. Deep neural networks ensemble for detecting medication mentions in tweets. J Am Med Inf Assoc JAMIA 2019;26:1618–26. doi:10.1093/jamia/ocz156.

[41] Wang X, Xu X, Tong W, Roberts R, Liu Z. InferBERT: a transformer-based causal inference framework for enhancing pharmacovigilance. Front Artif Intell 2021;4:659622. doi:10.3389/frai.2021.659622.

[42] Smith CL, Eppig JT. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med 2009;1:390–9. doi:10.1002/wsbm.44.

[43] https://ontox-project.eu/. [Accessed 25 February 2022]

[44] https://pubmed.ncbi.nlm.nih.gov/. [Accessed 7 February 2022]

[45] https://vhp4safety.nl/. [Accessed 25 February 2022]

[46] https://aopwiki.org/aops/18.