EXPLORING THE USE OF ARTIFICIAL INTELLIGENCE (AI) FOR DATA EXTRACTION AND VALIDATION FOR TOXICOLOGICAL RISK ASSESSMENT

Cao Yuewei

M.Sc. Thesis
February 2023

Student Number: 1053762
Course Code: TOX80436
Department: Toxicology (TOX)
Supervisor: Benthe van der Lugt

CONTENTS

List of tables 4

List of figures 5

List of Abbreviations 6

Abstract 7

CHAPTER 1. Introduction 8

- 1.1 Scientific perspective 8
  - 1.1.1 New approach methodology (NAM) 8
  - 1.1.2 Adverse outcome pathway (AOP) 9
  - 1.1.3 Tested chemicals 9
- 1.2 Technical perspective 10
  - 1.2.1 Artificial intelligence (AI) in toxicology 10
  - 1.2.2 Adverse outcome pathway (knowledge base) development 11
- 1.3 Workflow introduction 13
  - 1.3.1 Initial review 13
  - 1.3.2 Full-text review 14
  - 1.3.3 Data validation 15
- 1.4 Research question 15

CHAPTER 2. Initial review 17

- 2.1 Materials and methods 17
  - 2.1.1 Tool introduction 17
  - 2.1.2 Label creation 17
  - 2.1.3 Data collection 18
  - 2.1.4 Statistical analysis 20
- 2.2 Results 21
  - 2.2.1 Improvement of performance 21
  - 2.2.2 Distribution description 22
  - 2.2.3 Relevance between labels and classification result 23
- 2.3 Discussion 24
  - 2.3.1 Initial review 24

CHAPTER 3. Full-text review 26

- 3.1 Materials and methods 26

3.1.1 Tool introduction 26
3.1.2 Data collection 26
3.1.3 Model training 27
3.1.4 Statistical analysis 27
3.2 Results 28
3.3 Discussion 30
CHAPTER 4. Data validation 32
4.1 Materials and methods 32
4.1.1 Platform introduction 32
4.1.2 Word embedding and data validation 32
4.1.3 Parameter setting 33
4.1.4 Question generation methods 33
4.1.5 Data collection 34
4.2 Results 34
4.3 Discussion 35
CHAPTER 5. Tool improvement 37
5.1 Feasibility of using the GPT-3 model for data extraction 37
5.2 Methods 37
5.2.1 Create the prompt 37
5.2.2 Data Acquisition 38
5.2.3 Data analysis 39
5.3 Results 39
5.4 Discussion 41
CHAPTER 6. Conclusions and recommendations 44
6.1 Conclusions 44
6.2 Recommendations 44
Appendix A List of optimized search terms 46
Appendix B Figures of questions designed for testing the word embedding function 47
References 49

LIST OF TABLES

Table 1 The name of the labels and the corresponding questions 18
Table 2 Relevance between labels and the classification result 24
Table 3 Normalized extraction results 30

LIST OF FIGURES

Figure 1 Key steps of integrating big data into predictive toxicology from generation to application, related challenges and opportunities 11

Figure 2 The seven workflow steps designed to search, extract, harmonize and validate an AI-supported use of NAM data 13

Figure 3 Automatically generated histogram example 19

Figure 4 Changes in prediction scores during model training and update 22

Figure 5 Precision score at different thresholds after training rounds 22

Figure 6 F1 score at different thresholds after training rounds 23

Figure 7 Data collection and model training workflows in full-text review 27

Figure 8 F1 score with different labels, extraction methods and retraining methods 29

Figure 9 The workflows of prompt creation and evaluation in data extraction on GPT-3 model 38

Figure 10 GPT-3 performance in data extraction 40

Figure 11 Example output with an undefined label generated 41

LIST OF ABBREVIATIONS

Adverse outcome	AO
Adverse outcome pathway	AOP
Adverse outcome pathway knowledge base	AOP-KB
Artificial intelligence	AI
Bis(2-ethylhexyl)phthalate	DEHP
Butylbenzyl-phthalate	BBP
Deep neural network	DNN
Di-butylphthalate	DBP
Di-isononylphthalate	DINP
Diisodecylphthalate	DIDP
Key event	KE
Key Event Relationship	KER
Mode of action	MOA
Molecular initiating event	MIE
Natural language processing	NLP
New approach methodology	NAM
Organisation for Economic Co-operation and Development	OECD
PubMed ID	PMID
Quantitative structure activity relationship	QSAR
Weight of evidence	WoE

ABSTRACT

Lots of effort has been made to reduce, replace and refine the current animal studies. The new approach methodology (NAM) is an alternative, converting from direct observation to a less harmful but mechanism-focused method. The adverse outcome pathway (AOP) method is developed as a support to NAMs but requires a huge amount of manual work to conduct.

To automate the laborious process, artificial intelligence (AI) is applied. In this thesis, the suitability of the selected AI tools for data extraction and integration was evaluated in different case studies. Based on the user experience and the analysis of the tool performance data, the features of the tools were summarized and analyzed to predict their feasibility in practical application.

The tool used for screening the eligibility of publications in the stage of the initial review was Sysrev. It turned out that the tool performance improved during iterations and the platform update resulted in a noticeable improvement in the short term. But it was also indicated that the performance was influenced by other ongoing projects on the platform.

It can be seen from the full-text review, where key information for AOP mapping was extracted from publications, that different extraction methods result in different number of entities during the retaining of the model SpaCy. It was also shown some labels were more likely to be extracted from the full text than others, and this unbalanced extraction performance was the problem to be addressed in the tool improvement part.

During the validation of the extracted data, the natural language processing model call GPT-3 was applied to test its word embedding function. Questions related to the reliability of the publications were created. By challenging the model, it was found the model showed the capacity to recognize the semantic difference.

To address the problem in full-text review, the GPT-3 model was also applied in the full-text review. The result indicated that the GPT-3 model is a fast learner in the few-shot tasks, and that the model has in general a good performance on labels where the SpaCy had a poor extraction performance. It could also save the required manual work in entity sorting and showed the potential for quantitative AOP mapping. However, to put GPT-3 model into practical use, it should first overcome the limit on the number of tokens and improve the quality of the training data.

CHAPTER 1. INTRODUCTION

1.1 Scientific perspective

1.1.1 New approach methodology (NAM)

The current animal test is directly observing the adverse health effect *in vivo* to assess the risk. However, it has been recognized that the use of the animal test in chemical risk assessment has various limitations ranging from ethical problems to the lack of focus on the mechanism at a molecular level[1]. The principle of 3R (Replacement, Reduction and Refinement) was launched in the early 1960s to reduce the use of vertebrate animal testing, improving animal welfare and protecting human health. The 3R principle, as well as the concerns about extrapolating animal study data to human, leads to the movement from the animal tests to NAM.

The NAM currently has only working definitions, ranging from any approach, methodology or technology that can realize the goal of 3R, to the high throughput screening omics and especially in-silico computer modelling strategies[2]. Compared with animal experiments, the NAM has the advantage of exploring the chemical mode of action (MOA), extrapolating experimental data to human health as well as reducing the cost.

In practice, the NAM is adapted into five specific ways which include: A) the use of existing data, B) the use of weight-of-evidence approaches, C) information generated through quantitative structure activity relationships (QSARs), D) *in vitro* test methods and E) groupings of substances and read-across methods[2]. These five ways can be performed individually or in combination based on the acceptable uncertainty and the quality of available data[3].

In this thesis, the definition of NAM is restricted as followed to be consistent with the overall project. The NAM in this thesis includes all assays where the following models (or combinations thereof) were used:

- *in vitro*, *in chemico*, and *in silico* models, and/or;
- *ex vivo* models of all vertebrates, to which the chemical of interest was NOT exposed during lifetime, but directly on isolated tissues, and/or;
- 3R whole body organisms (e.g. zebrafish embryos (max. 72 hours post-fertilization), C. elegans and Drosophila melanogaster).

1.1.2 Adverse outcome pathway (AOP)

The AOP converts the traditional toxicological tests to a more mechanistic and data-driven way[4], allowing a more efficient grouping and prioritizing of chemicals. The combination of AOP and other NAM methods such as read-across, the chemical assessment will be more related to the structural properties and thus reduce the use of animal testing[5].

The AOP provides a mechanistic representation of key toxicological effects covering the molecular, cellular, tissue, organ, organism and population levels[6]. The AOP map always begins with the molecular initiating event (MIE), progressing through a series of intermediate key events (KEs) which finally leads to the adverse outcome (AO). The relationship between key elements (Key Event Relationship, KER) defines the causal relationship between KEs and can be qualitative or quantitative. To develop an AOP, one can start from an MIE, KE or AO[6]. This process leads to a decomposition of a complex toxicological process into discrete and easily processed information. And if quantitative KER information is available, together with the understanding of the additional modulating factors (i.e. extrinsic factors not driven by the KE in an AOP)[7], a quantitative estimation of the influence from KE upstream to KE downstream (or AO) can be made which supports the quantitative chemical risk assessment[8].

Although the AOP can be used to reveal the mechanism of a certain chemical, it is not stressor specific. This means that it focuses on the dynamics but not kinetics, and it also suggests that a recognized biological pathway can be re-used in other AOP processes. Therefore, a single AOP can be seen as a part of a larger AOP network[9]. The AOP network provides insights into the adverse biological pathways and can be supportive in several aspects, including the testing and grouping of chemicals, the selection of methodologies for developing the guideline for chemical testing, and the interpretation of results from non-guideline testing methods[10].

1.1.3 Tested chemicals

The modelling of AOP is supported by a large and growing number of digital resources, which include models/databases/tools that are generated numerically or converted into digital formats[11]. The AI4NAMS project, initiated by EFSA, aims to facilitate this process by exploring the potentials of state-of-the-art AI tools and methods in searching, extracting, and integrating NAMs-based data[12]. However, it is not realistic to test the extraction and integration performance of the AI tools on all kinds of compounds. Instead, compound classes that are typical with regard to their kinetic properties and their mode of action[1] should be selected and further investigated. Aside from the typicality of the tested compound, it is most reasonable to include chemicals that have ample NAM data and provides a specific scenario for data extraction.

EFSA has indicated that minimally twelve case studies will be proposed including dioxin-like compounds, phthalates and pesticides, and 6 proposals based on endpoints including endocrine activity (related to EATS) and genotoxicity. In this thesis, phthalates will be the main focus with the support of the experience of other projects (e.g. the dioxins). The phthalates are commonly used as a plasticizer, but they can migrate from food contact plastic into food. Studies have shown that they can lead to the endocrine system disruption, reproductive tract abnormalities, cognitive impairment[13]. The dioxins are known as one of the carcinogens to humans, especially fetuses and infants. They can also lead to noncancerous effects in the nervous, urinary and endocrine systems[14].

1.2 Technical perspective

1.2.1 Artificial intelligence (AI) in toxicology

Because of the explosively growing number and diversity of toxicological and chemical data, the *in-silico* method is reckoned as a promising way for data selection, extraction and integration. The establishment of an AOP depends a lot on mining and integrating toxicological data from various resources, which would be laborious and low efficiency if it is manually done. As such, the AI is exploited to automate these processes.

The AI provides a new approach to extracting, processing and analysing a large amount of information. In most cases, the information of a biological process appears in a document in the form of free text, which is the so-called unstructured text/data[5]. By sorting the structured or unstructured toxicological data into databases such as AOP, these data can be validated and be used for further risk assessment.

The AI-predicted toxicity can be an indicator for prioritizing the chemicals to be assessed. Based on such predictive ability, the analysis results can be a valuable reference or support for toxicologists to find out the problems that are less noticed. Furthermore, the integration of various data stimulates the generation of new knowledge. For example, “AI could assist in identifying emerging risks, i.e. new adverse effects not typically covered by the standard regulatory endpoints[15].” Based on the explanation above, the process, potential opportunities and problems during integrating AI into predictive toxicology are concluded in Figure 1.

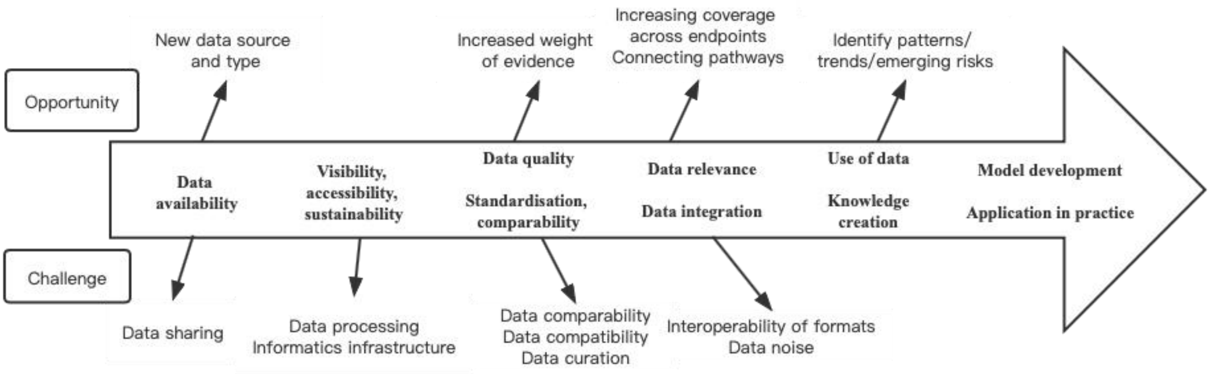


Figure 1 Key steps of integrating big data into predictive toxicology from generation to application, related challenges and opportunities (Developed from [16])

While the application of AI in toxicological risk assessment is promising, there are still gaps between the data and the AI performance. The application of AI still needs the involvement of experts. One reason is that without a robust evidence surveillance, the result may come from studies with reliability, relevancy or adequacy problems. Either significant evidence gaps or evidence gluts can lead to poor performance in data fitting and the results that are inconclusive or irrelevant. Therefore, it is necessary to have a careful check conducted by the experts on the data before inputting. In this case, the benchmark data manually added by the expert will be generated for every workflow step and will serve as surveillance during the process. The other reason is that the predictive result should be re-evaluated to see if the latent patterns the AI built is reasonable or not. The opinions of expert team also play an important role in this part.

1.2.2 Adverse outcome pathway (knowledge base) development

As is concluded by Spinu et al.[8], the development of a quantitative AOP requires information with regard to some critical aspects, which includes the modulating factors, the possibility for available data to be used for quantification, defining the threshold that causes an effect, mathematical modelling on KER and dose-response relationships, and model validation and implementation.

The KE and KER are the two fundamental units in developing an AOP. To identify them, the high-throughput and high-content data can be exploited in the early stage, which can be acquired by data mining. In practice, to integrate a KE or KER into an AOP map, special attention should also be paid respectively.

Each KE represents an essential measurable change in a biological process but not necessarily provide sufficient information[17]. The identification of a KE involves judgment about the relevance and abundance of the KE data. How many KEs and at what level should a KE be included have to be based on the opinions of experts[18]. The expert needs to determine how to find a balance between the availability and the reusability of the KE information included. Besides, the description of a KE also matters. Considering in most cases, people who are in charge of computational modelling may not be familiar with the biological process, the description of a KE should be legible enough to give guidance. Detailed information about the process and the measurement should be given to suggest the type of available data and the potential modelling approaches, mathematical, formalisms, and parameters that can be used during the integration[19].

The main concern in determining a KER is the validation of the causal nature between two KEs. The evaluation of weight of evidence (WoE) justifies the biological plausibility of the connection, i.e. the perturbation from the upstream event leads to the downstream event[18]. The Organisation for Economic Co-operation and Development (OECD) suggests the adoption of tailored Bradford-Hill (BH) considerations to assess the causality of associations[20].

To implement the weight-of-evidence procedure and multiple levels of peer review, an easily-accessed and integrated platform for others to obtain up-to-date AOP information will be of great help[21]. The adverse outcome pathway knowledge base (AOP-KB) is a database that facilitates the storing and querying of AOPs and KEs. In the real case, the establishment strategy of an AOP network is largely dependent on the feasibility of current toxicological data and still has some other limitations.

General toxicological information or statistical inference can only be applied in the initial steps, while more precise and systematic data will be further required in the steps of KEs and KERs definition[6]. The scientific literature is one of the important sources to provide such rigorous data, and prospectively generated data can also be used in the case of knowledge or data gaps[6]. However, the lack of fit-for-purpose data and the hardness of data generation lead to the barrier to a quantified AOP or low confidence in the evidence.

Apart from the lack of data, the characteristic of an AOP can still be an obstacle[22]. The typical AOP is a unidirectional linear map. But if the AOP bifurcates, or another pathway joins the AOP, the linearity will no longer exist. The concurrent KEs can be addressed by adding several temporal sequences but will be contrasted against the independence of a KE. In a complex biological process, the positive/negative feedback loops may be involved, making the quantification hard to perform. A solution is still lacking in such unidirectional linearity. Even in a linear AOP, there is still the possibility that the downstream KE is not only regulated by the upstream KE but other modulating factors, making the quantitative step more complex.

The case studies selected for further investigation in this project involve these obstacles above. By addressing the problems in the case study, experience will be gained to further improve AI performance in extraction and integration.

1.3 Workflow introduction

The introduction of AI into chemical risk assessment is a promising new approach, with great savings in human and material resources. By selecting the suitable tool and assuring the quality of data for use, the potential difficulties can be reduced to some extent. To realize this goal, the entire process of this project is depicted as Figure 2. The box in red shows the steps that this thesis will be focused on.

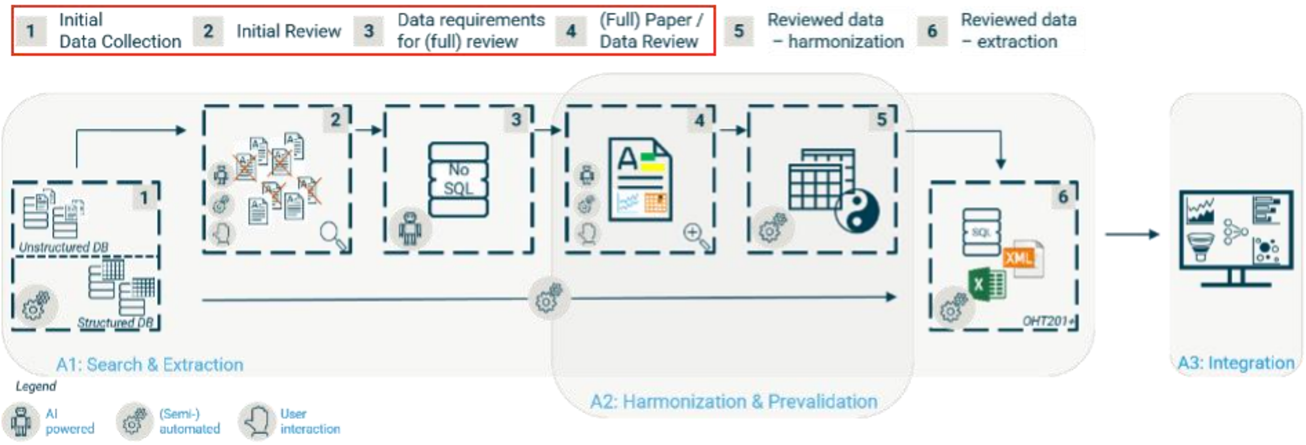

1 Initial Data Collection
2 Initial Review
3 Data requirements for (full) review
4 (Full) Paper / Data Review
5 Reviewed data – harmonization
6 Reviewed data – extraction
Unstructured DB
Structured DB
No SQL
SQL
XML
OHT201
A1: Search & Extraction
A2: Harmonization & Prevalidation
A3: Integration
Legend
AI powered
(Semi-) automated
User interaction

Figure 2 The seven workflow steps designed to search, extract, harmonize and validate an AI-supported use of NAM data[12]

1.3.1 Initial review

The collection of fit-for-purpose data is always the first step in building or improving an AI tool. The extraction of fit-for-purpose publications was performed in the initial review. However, considering the data availability, the two-step extraction was performed separately on different chemical case studies.

During the initial review stage, data will first be collected from structured and unstructured databases (e.g. ToxCast and PubChem) with specific keywords. Then the free text will be imported into Sysrev, a data mining tool for systematic search and extraction of data from documents. The free text, including the title and the abstract of a publication, is seen as unstructured data in the field of machine learning. These unstructured data will serve as the input to test the Sysrev's inclusion/exclusion functionalities. An xlsx file will list all references with the result of inclusion/exclusion. The Sysrev machine learning model predicts the eligibility of the document to be selected for further use and outputs a list with entries for each document and the prediction score. Based on the score, the fit-for-use document will be selected and further processed. A possible risk for this would be that the amount of retrieved data is too low to continue to the next workflow step.

1.3.2 Full-text review

Before the full-text review, the data obtained in the initial review should first be standardized and stored as the full-text information. The parsed PDF file after the initial review was stored and formed into a JSON object by the database called Elasticsearch. Considering the possibility of a lack of data, publications by purchase or from other data sources can be added to the initially reviewed publications in this stage. By querying this database Elasticsearch, a JSON object will be returned, containing specific parts (e.g. title, author, abstract, full text, etc.) of one or more publication(s) stored in the index.

A full-text review is performed for the methodology (e.g. test system and test material information) and result (i.e. Process, Object, Action) extraction. The returned JSON object will be processed as the input for the full-text review. During the full-text review, the applicable text/PDF annotation tool (e.g. PDFHighlighter) and the natural language processing (NLP) models from medspaCy will be applied. The annotation tool is used to visualize the process for an easier inspection and adjustment. The medspaCy provides packages that are necessary for processing common clinical text specific to the clinical domain. Experiments will be performed to test the labelling applicability of the medspaCy-based packages with regard to the NAM-focused publications.

The result of the full-text review will be mapped to the OECD harmonised template no. 201 (OHT 201), which is designed to store and categorize the data on intermediate effects and mechanistic[23]. The expected delivery should include the criteria on the benchmark data label, benchmark data, the manually selected and AI-performed annotation of the text, suggestions on the tool improvement and an evaluation on completeness and coverage between the expert and AI-based approach.

1.3.3 Data validation

Before any harmonization or integration is done, it is essential to first verify the key information which will be the basis of the AOP mapping. This step is essential to gain the universal regulatory acceptance and public confidence[24].

According to the conclusion by Van der Zalm et al.[24], the validation of key information for AOP endpoints includes five essential elements: fitness for purpose, human biological relevance, technical characterization, data integrity and transparency, and independent review. These criteria cover a wide range of specific terms that help build up the scientific confidence. In the meantime, they can still be adjusted with regard to e.g. the order of evaluation, the level of acceptance according to the characteristics of the publications in practical use.

In this section, the PGT-3 model will be used to validate the reliability. The GPT-3 model is a promising tool to process natural language. The specific function of the GPT-3 model that will be exploited is word embedding. The word embedding is an algorithm to encode and decode the natural language so that words with similar meanings can be grouped. By using the word embedding function, the model is expected to predict the expected word with high accuracy. Besides, according to the output of the model, the quality of these questions will be evaluated and adjusted.

1.4 Research question

In this thesis, the suitability of the selected AI tools for data extraction and integration will be evaluated in different case studies. Based on the user experience and the analysis of the tool performance data, the features of the tools will be summarized and analyzed to predict their impacts on the practical application of the tools. For further improvement, the information about tool performance, usage experience and tool improvement will also be considered to find the potential improvement methods. The possible improvement will be performed on a database with a smaller scale and get evaluated.

To realize these goals, the performance data during the initial review and full-text review will be analyzed using metrics that evaluate the classification capacity. Based on these metrics, statistical analysis will be used to further find out the latent relationship within the data and suggestions for tool improvement will be given. Since the outputs of the full-text review were quite complex, a visualization step will be first performed to help with the data analysis. The same analysis steps will be followed and improvement suggestions will be given. In the validation step, the word embedding function will be tested by evaluating the reasonableness of the output, which is the answer to some manually created questions about the potential risk of bias.

CHAPTER 2. INITIAL REVIEW

2.1 Materials and methods

2.1.1 Tool introduction

To extract and label relevant publications from a variety of data sources, the platform Sysrev was exploited. Based on the principle of FAIR (Findability, Accessibility, Interoperability, and Reuse of digital assets), Sysrev is a data mining tool that provides a public access to systematically review the unstructured scientific publications regardless of the disciplines[25]. Data management on Sysrev generally follows 4 steps:

A) Collecting articles from specific sources;
B) Defining labels as the selection criteria;
C) Reviewers recruitment and review tasks implement;
D) Analysis of exported data.

Sysrev supports the import of data in the form of XML, RIS and PDF. It also provides an API to medical publication databases such as PubMed. In this stage, the API for direct search in PubMed was used for the initial search. The PubMed IDs (PMIDs) of the resulting publications were saved as a text file and imported into Sysrev.

The labels for publication review are created by the users and serve as the criteria for classification. The platform periodically updates the machine learning model with results from all projects performed on the platform. The model automatically generates histograms describing the model's prediction accuracy for manually labelled and categorized publications, which also serve as the data for model training[25]. Although Sysrev's machine learning model has already been trained on other projects, to better fit it to the phthalate case, a manual review is still needed.

2.1.2 Label creation

To accurately assess the inclusion or exclusion of the publications, criteria were defined and added as labels in Sysrev. The Boolean (logical judgment with only true/false values) labels were created. For each publication, the viewer will make a judgement according to each of these labels.

The specific definition of these labels was discussed in advance to reduce interpersonal variation. The labels and the corresponding questions to elaborate the labels are displayed in Table 1. The interpretation of the label 'NAM' is solely based on the definition in this project (see 1.1.1). By making a judgement on this label, publications that don't belong to the NAM study were excluded. The 'chemical of interest' label was used to exclude publications that mention the chemical in a different context. Publications where mixtures of chemicals were tested, were also included. The label 'Human relevance' guarantees the exclusion of endpoints related to e.g., non-test model animals and environmental toxicology. Only the original experimental research papers were considered, thereby excluding reviews and epidemiological studies, amongst others[26]. The label 'Include' was definitive to the classification result and was based on the judgements on the previous labels. A positive answer was only given to this label when all the former labels were answered with "Yes".

Table 1 The name of the labels and the corresponding questions

Labels	Corresponding questions
NAM	"Does the study include (partly) NAM data?"
Chemical of interest	"Is the chemical of interest tested in the right context?"
Human relevance	"Is the tested endpoint relevant for human health?"
Toxicological relevance	"Is the endpoint of the paper relevant from the toxicological perspective?"
Experimental paper	"Is the publication an original experimental paper?"

2.1.3 Data collection

The initial review data was obtained from the phthalates case. In this case study, specific kinds of phthalates and their metabolites were focused on, including di-butylphthalate (DBP), butylbenzyl-phthalate (BBP), bis(2-ethylhexyl)phthalate (DEHP), di-isononylphthalate (DINP) and diisodecylphthalate (DIDP), as these are authorized for use in food contact materials in the EU[26]. According to the aim of this project, data should come from scientific publications that are: exploiting the NAM method (or partly, e.g. both *in vivo* and *in vitro*); testing the chemicals of interest, i.e. phthalates or its metabolites for initial review including DBP, BBP, DEHP, DINP, DIDP and seven high-potency polychlorinated dibenzodioxins for full-text review; original experimental research papers; with human-relevant endpoint; with the endpoint that is relevant from the toxicological perspective.

Before this thesis, publications had been searched directly in the PubMed database using the keywords of method element and chemical element (Table 1 in Appendix A). The benchmark data to evaluate the model training was collected for DBP, as it was discovered in the previous phase of the project that this was one of the phthalate types with the highest data availability[27]. Based on the comparison between the benchmark data and the machine prediction, the metrics evaluating the tool’s performance can be generated and collected. These publications were imported into Sysrev serving as the database for the initial review.

In total, 150 publications were selected as the benchmark data and imported into Sysrev. These publications were manually reviewed by 2 experts to avoid potential bias. The decisions on including or excluding a publication were based on information present in the title and abstract with regard to the labels in Table 1. The review process resulted in 77 eligible publications meeting the inclusion criteria. 68 publications were excluded, and 5 publications were considered as ‘unsure’, and therefore neither included nor excluded.

During the expert review, randomly selected publications from the Sysrev database were first manually reviewed by the first reviewer. A group of 30 included publications and 30 excluded publications was seen as a set and reviewed by a second reviewer. The sets of publications, including the title and abstract of these publications, were processed during the periodical update to train the Sysrev machine learning model. Every time Sysrev was updated using the data from all projects on the platform, the result data were exported to the evaluation form and the automatically generated histogram was also screenshot. An example of the histogram can be seen in Figure 3. The X-axis indicates Sysrev’s confidence that a publication should be included. The green bars are publications that were manually included, and the red bars are publications that were manually excluded. In an ideal case where the machine predicts with high accuracy, the green bars should be concentrated on the right side of the histogram and the red bars on the opposite side.

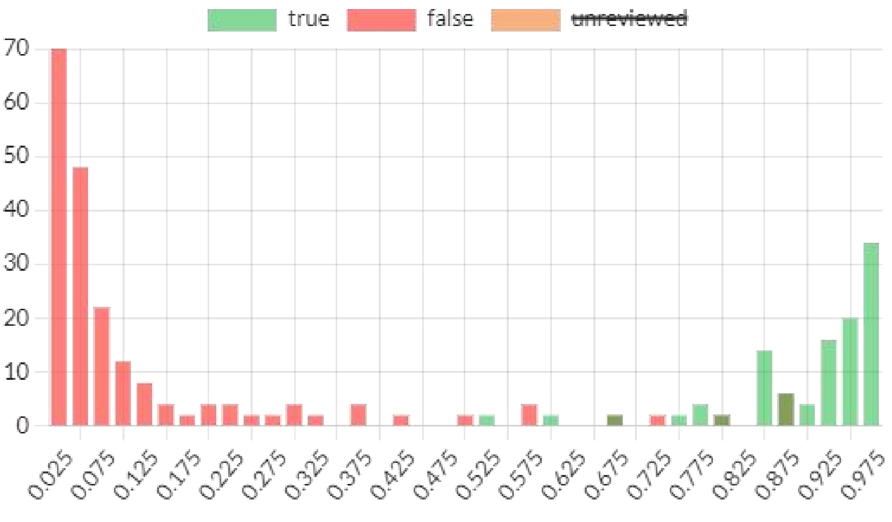


Figure 3 Automatically generated histogram example

To extract data from these histograms, a tool called WebPlotDigitizer was applied. It was found that the WebPlotDigitizer often failed to identify the bars where the colours overlayed. Therefore, a manual extraction mode[28] was adapted by manually adding dots on the top of each bar. A default value in the classification task to classify the inclusive and exclusive publications was set as 0.5. In this case, the red bars with a horizontal coordinate value less than 0.5 were considered as the number of publications correctly excluded, while red bars with a horizontal coordinate value more than 0.5 were considered as the number of articles incorrectly excluded. And as for the green bars, the value 0.5 was also the threshold to classify the inclusive and exclusive publications.

2.1.4 Statistical analysis

To evaluate the performance of these two tools, the three metrics, i.e. recall, precision and F1 score, were applied. These three metrics are the indicators of the classification capacity on different dimensions. The precision score, calculated as $\frac{\text{True Positive}}{\text{Ture Positive + False Positive}}$, is the indicator of the purity of the result. The use of precision score indicates the possibility that a data point is correctly classified from the irrelevant data points. The recall score is calculated as $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$ and it indicates the efficiency of the classification or in other words, how much true positive data is not wrongly discarded during the process. Since the precision score and recall score focus on Type I and Type II error respectively, it is not difficult to understand there is a tradeoff between these two metrics[29]. To evaluate the overall performance of the tools, considering both the purity and efficiency of the search result, the F1 score can be applied. The F1 score is defined as the harmonic mean of the precision score and the recall score, i.e. $\frac{2}{\text{Precision}^{-1}+\text{Recall}^{-1}}$.

Based on the different aims of initial review and full-text review, different metrics should be exploited according to their focus. During the initial review, more emphasis is on the purity of the suitable publications so that the precision score will be the main standard. While during the full-text review, the aim is to extract all the possible AOP-related subjects. Therefore, not only the purity but also the efficiency of the data extraction should be evaluated using the F1 score.

In the initial review, it is assumed by default that only with a possibility over 0.5 can a positive/negative prediction be deemed as positive/negative. However, there is a possibility that in some cases the data follow an extremely uneven distribution, the default threshold of

0.5 cannot represent the data features properly. Therefore, aside from the classification analysis at the defaulted threshold, an evaluation of changes in prediction scores at different thresholds over time was also performed.

To further analyze the differences between labels, the one-way analysis of variance (ANOVA) test was used in SPSS. A value of $p<0.05$ was assumed as statistically significant. A Chi-Square test, followed by the calculation of Cramer's V value, was used to quantitatively analyze the strength of an association between labels and the initial review result.

Aside from the F1 score which indicates the quality of the extraction, the number of entities extracted under each label was also taken into account. By multiplying the number of entities and the corresponding F1 score, the result represents a combination of extraction efficiency and extraction accuracy. A K-means cluster analysis was then performed on this result to rank the extraction performance in SPSS.

2.2 Results

2.2.1 Improvement of performance

In the stage of initial review, the precision score should be reckoned as the main standard to focus on correctly classified publications. But to ensure that the prediction score does not go to the extreme, an F1 score was also checked as a reference. The figure below (Figure 4) displays the changes in the precision score and the F1 score after each evaluation run to reflect the Sysrev performance. By inserting the average value of all precision scores as a dashed line in between, the changes before and after the platform update can be better perceived. The difference in the background colour is due to the fact that there was a major machine learning model update between the third and the fourth evaluation run.

During the training, the prediction score trended upward but not continuously increased. Comparing the prediction scores before and after the model update, the precision score and the F1 score increased to 1.4 and 1.3 times respectively, suggesting that the model performance got a major improvement due to the update. However, the impact of the update did not last long because there is no significant difference between the first three and the last three runs ($p>0.05$).

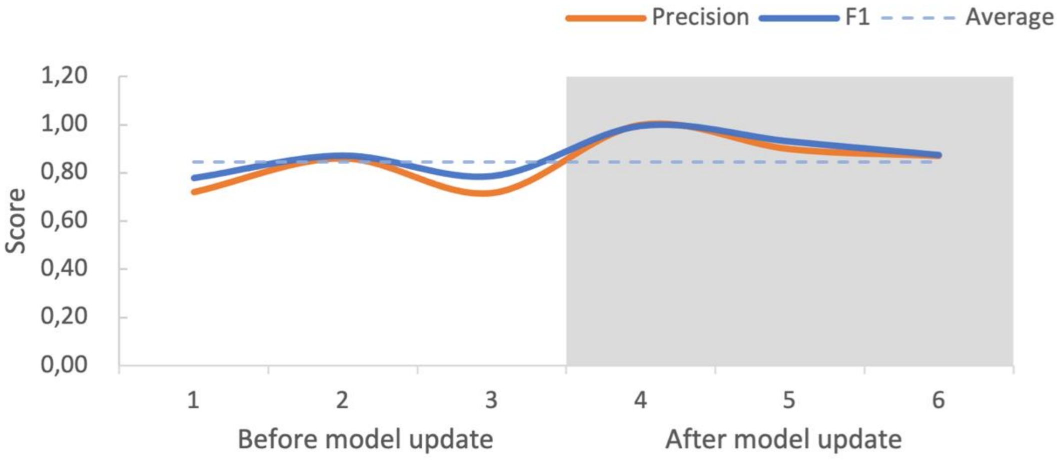


Figure 4 Changes in prediction scores during model training and update

2.2.2 Distribution description

In the method part, it was pre-assumed that the threshold for classifying the inclusive and the exclusive publications is 0.5. It facilitates the analysis but neglects the information about the prediction with different possibilities.

In order to reveal the overall distribution of predicted data after each evaluation run, the threshold of classification was set at different values, separating the positive prediction and the negative prediction differently. The performance data at different thresholds (0.1, 0.3, 0.5, 0.7) was collected on both precision score and F1 score after each evaluation run (Figure 5, Figure 6).

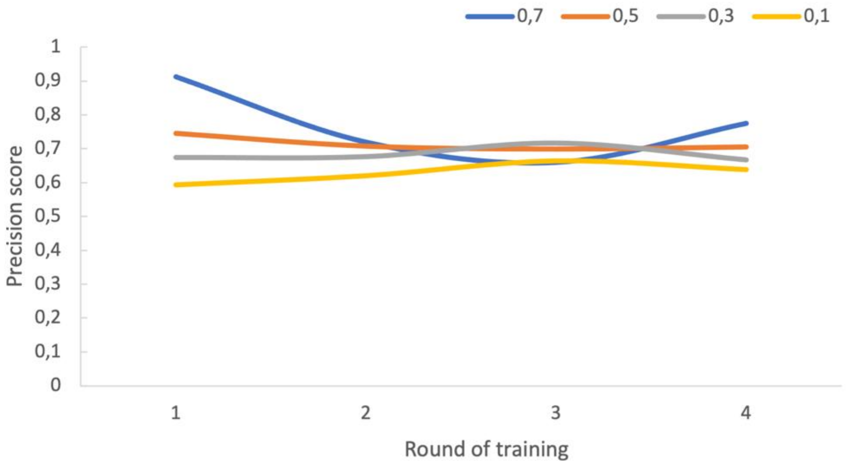


Figure 5 Precision score at different thresholds after training rounds

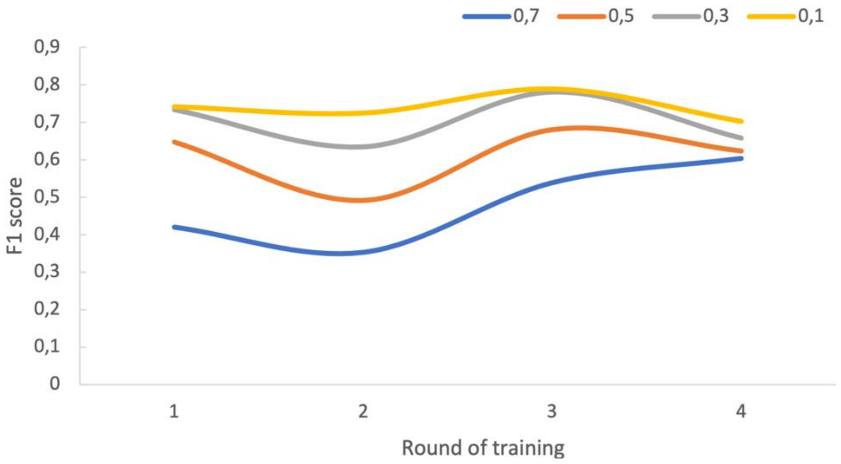


Figure 6 F1 score at different thresholds after training rounds

Based on the analysis of the precision score at different thresholds, only the curve representing the inclusive threshold of 0.7 showed an obvious U shape during the training, while the other curves were comparatively flat. And the average value of these curves concentrated around 0.7 without significant difference ($p>0.05$). This indicates that with a higher inclusive threshold, the classification of negative cases did not change much. And it can also be deduced that the value of 0.5 as the threshold can represent the tool performance of most of the data with regard to precision.

Compared with the precision score, the F1 score showed a clear difference in values between different thresholds, which should be attributed to the difference in the recall value. A negative association can be found between the value of the threshold and the F1 score. Based on the calculation process of the F1 score and the fact that the precision score did not change much, it can be deduced that the recall score decreased during the training, i.e. the number of false negative publications increased. Combining the two scores that have been discussed, it can be concluded that the changes in tool performance were mostly due to a stricter classification of the publications that should have been included.

2.2.3 Relevance between labels and classification result

An understanding of the inclusive publications may provide useful information about tool improvement. Therefore, the relationship between labels and the prediction of classification was also examined (Table 2).

Table 2 Relevance between labels and the classification result

Labels for correlation	Cramer's V	Pearson Chi-Square
Chemical of interest*Result	0.749	0.00*
Toxicological relevance*Result	0.364	0.00*
NAM*Result	0.299	0.00*
Human relevance*Result	0.229	0.004*
Experimental paper*Result	0.061	0.44

A Pearson correlation coefficient of less than 0.05 indicates there is a significant association between the label and the result. And the Cramer's V value indicates how strong the association is between two objects, serving as strong evidence for the Chi-square test results. In most cases, the Cramer's V value is considered weak when it is less than 0.1, moderate when in the range of 0.1-0.5 and strong when more than 0.5[30].

Based on this, the label 'Chemical of interest' is strongly related to the in/exclusive result while the 'Toxicological relevance', 'NAM' and 'Human relevance' are moderately related. The label 'Experimental paper' has weak and insignificant relation with the in/exclusive result. Considering the selective criteria, i.e. one document should be excluded once any one of the inclusion criteria is not met, a stronger relationship actually indicates a higher chance of rejection due to the failure to meet the criteria. In this way, the interpretation of the high relevance between the label 'Chemical of interest' and the inclusion result should be that most of the documents were excluded because the phthalates are not the object of study.

2.3 Discussion

2.3.1 Initial review

Due to the automatic update and the generation of histograms, Sysrev is user-friendly with regard to the model training and result collection processes. It was shown that the model update resulted in larger improvement than the mere training. The improvement due to the update was probably because the data used for this update mainly came from the phthalates project, making the updated model fitter for this project.

Based on the information available, the variance of the Sysrev model is unknown. However, since it's quite common in machine learning models to have varying outputs, the improvement during training can also be attributed to this. The variance in the Sysrev classification performance can possibly be due to the inherent variance of a machine learning model, expert behaviour within the group and the external influence from other projects.

The inherent variance is attributed to the uncertainty in the machine learning process. The data for model training is always only a part of the entire problem domain, therefore there exists some difference from reality. In most cases, the algorithm of a model is deliberately added with some random numbers, so that the bias in the sample could be better overcome. The random number in the algorithm leads to a varying output but within an acceptable range. The variance can also come from the discordance between reviewers because the conflicting classification judgement can also be used during the model training. Such kind of conflicting data can be seen as the noise in the input, which adds uncertainty to the performance of the model. Besides, as is shown in the prediction scores, the model can be easily influenced by the ongoing projects performed in the platform. In other words, the judgement from other projects also influenced the output of the phthalates case. All these factors influence the overall stability of the output. The variance would lead to a varying predicted distribution, which can be seen from the analysis of the classification threshold, leading to a loss of usable data. To reduce the inevitable variance in the model output, the application of Sysrev requires a high amount of training data with high quality so that the noise in one project can be neglected or amended by the data from other projects.

Another possible method to improve the tool performance would be improving the label-specific classification performance. The relevance between labels and the classification result suggests there is an imbalanced distribution in the training data with regard to this label. In this way, relatively more information about the small class would be useful to improve the performance of extraction[31]. This could be realized by optimizing the search method in the database to get a more balanced data set. For instance, since most of the publications were discarded because of being irrelevant to the ‘Chemical of interest’, it would be a solution that when searching the PubMed database, enlarge the chemical-related keywords to increase the percentage of relevant publications among the input publications.

CHAPTER 3. FULL-TEXT REVIEW

3.1 Materials and methods

3.1.1 Tool introduction

In the dioxins case, the following tools were tested:

SpaCy (“en_ner_bionlp13cg_md”)
PDFHighlighter tool
Fitz (PyMuPDF)
medSpaCy

Based on the feedback of the dioxins case, the spaCy model (“en_ner_bionlp13cg_md”) was re-trained and tested. It is an NLP model for processing biomedical, scientific or clinical text, which can be used to build information extraction or natural language understanding systems or to pre-process text for deep learning.

The PDFHighlighter was used as an online tool, to manually label entities from full text publications. Fitz was used to translate the manually labelled entities from the highlighter to their position within the PDF text. Furthermore, the sentence extractor of SpaCy’s model “en_core_web_sm” was used, to map the entities to the sentences where they appear, to ensure the right data format. This a way of extraction is called the highlighting method.

As another way of extracting data, medSpaCy’s functionalities were leveraged to automatically label entities from PDF texts. Such a way of extraction is called the distant supervision method.

3.1.2 Data collection

The full-text review was performed on full texts of all references retrieved and selected in the initial review. In the highlighting method, the data used for full-text review was obtained by labeling the content of selected publications using the PDFHighlighter. The labels were created in advance according to the entities of interest in the OHT201. Only OHT201 priority fields were labelled, namely test system identity, type of test system, substance, process, object, action, parameter, result for the parameter and controls / reference items[32]. In total, 40 publications were manually labelled within the results and methods sections of these publications. The extracted data were used to retrain and test the tool SpaCy.

Different sources of data were used in the distant supervision method. The data either came from the manually extracted labels or benchmark publications which were not labelled manually from the previous case study. Similar to the highlighting method, only the results and methods sections were used.

3.1.3 Model training

The manually labelled data was randomly split with a ratio of 6:2:2 into a training, validation and test set. The training and validation sets were used to retrain the models, while the test set was used to benchmark the retrained models as well as the pre-trained "en_ner_bionlp13cg_md" model.

As a summary of this part, the data collection and model training methods used in this chapter can be represented in the diagram below.

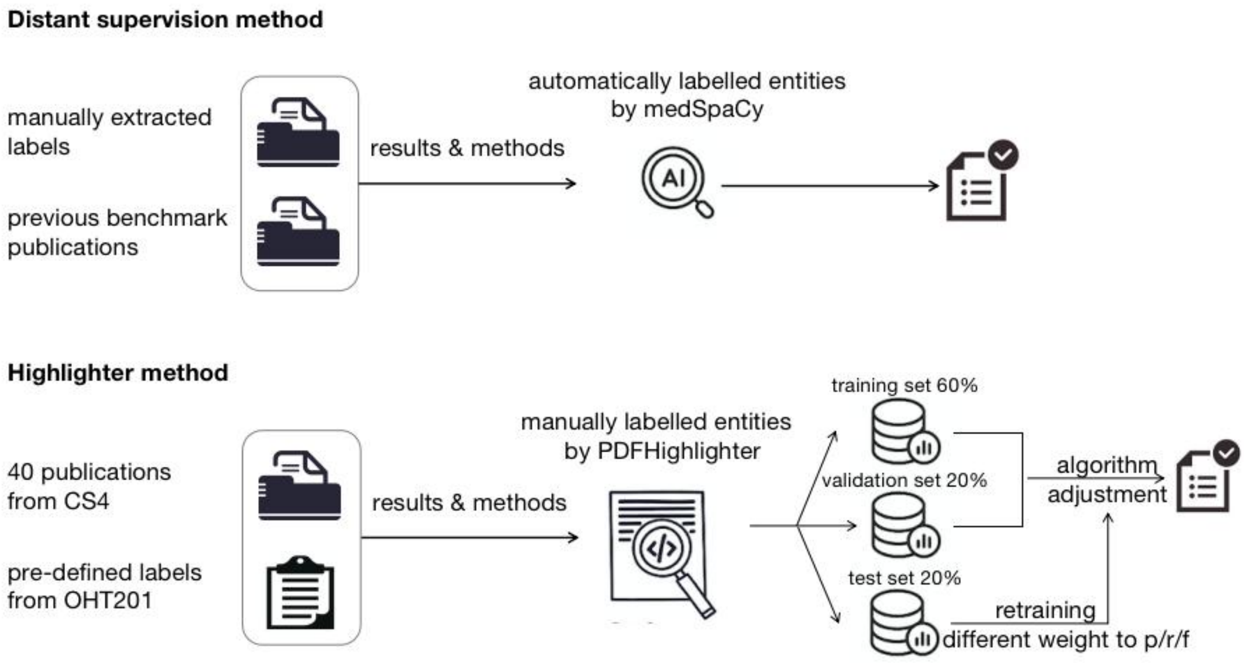

Distant supervision method
manually extracted labels
previous benchmark publications
results & methods
automatically labelled entities by medSpaCy
AI
Highlighter method
40 publications from CS4
pre-defined labels from OHT201
results & methods
manually labelled entities by PDFHighlighter
training set 60%
validation set 20%
test set 20%
algorithm adjustment
retraining
different weight to p/r/f

Figure 7 Data collection and model training workflows in full-text review

3.1.4 Statistical analysis

The precision, recall and F1 scores were also calculated in this step to evaluate the label-specific extraction performance. Different from the previous step, the full-text review aims to extract all the possible AOP-related subjects from the publications. Therefore, not only the purity but also the efficiency of the data extraction should be evaluated, i.e. the F1 score should be the main metric in this stage. To visualize the differences in F1 scores with different labels, extraction methods and retraining methods, a heatmap was generated using the OmicStudio tools at https://www.omicstudio.cn/tool.

Aside from the F1 score which indicates the quality of the extraction, the number of entities extracted under each label was also taken into account as a sign of the efficiency of the extraction. By multiplying the number of entities and the corresponding F1 score, the result represents a combination of extraction efficiency and extraction accuracy. A K-means cluster analysis was then performed on these results to rank the extraction performance in SPSS.

3.2 Results

The heatmap visualization reflects the prediction score variation between different data labels, extraction methods and retraining methods. A larger value is associated with a more intense red colour.

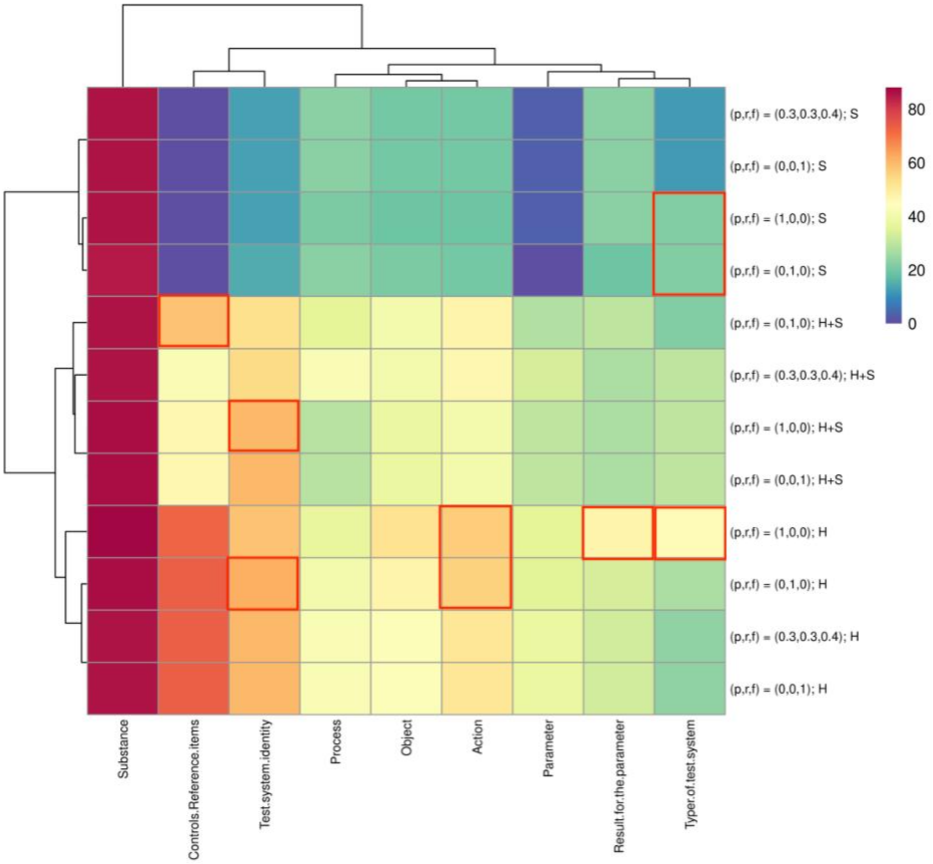


Figure 8 F1 score with different labels, extraction methods and retraining methods

The heatmap above depicts the prediction scores distribution over labels, grouped by the combination of the extraction methods (supervisor, highlighter, supervisor+highlighter) and the retraining methods (different weight to precision, recall and F1 score). From the heat map, it can be seen from the most intense red colourings in the left-most column that the highest scores were always associated with the Substance extraction. And the scores of control reference items and test system identity showed obvious variance between extraction methods. The pattern in cell colours across extraction methods also suggests that the supervisor method (the first 4 rows) generally resulted in lower scores compared with highlighter and supervisor+highlighter methods.

The heatmap also supports the result of the ANOVA test. Seeing the difference between columns, the label 'Substance' showed a significantly higher score than other labels no matter what the extraction methods and training methods were. It can be found from the graph that there is a significant difference in colour between these groups, which means different extraction methods led to a significant difference ($p<0.05$) in the prediction scores. When comparing the difference within the group of every 4 rows, in general, no colour difference was observed. This was confirmed by the ANOVA test that giving different weights to precision/recall/F1 score to retrain the model does not significantly influence the prediction score ($p>0.05$). But a higher extraction performance was more likely to occur when a higher weight was given to the precision/recall score alone (e.g. cells boxed in red). This is perhaps linked to the unbalanced distribution of the corresponding content in the publications.

What is not shown in this graph is the number of extracted entities. Since in practice, the data for model training should be ample and accurate, this factor cannot be neglected.

Based on the absolute value of the number of entities, different extraction methods (supervision/highlighter/supervision+highlighter) result in different number of entities under labels however no extraction method has the highest number of entities consistently. And because of the lack of labelled data in publications in the distant supervision method, the label of controls reference items had no entities extracted.

Combining the number of entities and the prediction score would be a more comprehensive way to evaluate the tool's performance, as the number of entities shows the efficiency of the tool in finding and extracting keywords while the prediction score shows how accurately the tool predicts.

Table 3 Normalized extraction results

Labels	Extraction methods		
	Highlighter data	Supervisor data	H+S
SUBSTANCE	100.00	100.00	33.71
OBJECT	37.47	19.59	53.94
TEST SYSTEM IDENTITY	38.79	0.89	34.96
ACTION	34.13	10.91	100.00
PROCESS	21.15	5.94	45.44
PARAMETER	13.22	0.36	28.59
RESULT FOR THE PARAMETER	7.19	5.95	8.77
CONTROLS REFERENCE ITEMS	11.85	0.00	6.17
TYPE OF TEST SYSTEM	2.28	9.09	79.68

* The cluster analysis result is shown in colour. A darker colour represents a higher class of value.

According to the result of cluster analysis, different extraction methods had label-related performance in keywords extraction. In general, the labels 'Action', 'Object', 'Substance', 'Process' and 'Test system identity' showed a comparatively better result. And by contrast, some labels were always at the bottom. Labels that are related to the biological process, including 'Action', 'Object' and 'Process', in most cases, can be recognized and extracted well. In general, the scores for highlighter data are significantly ($p<0.05$) higher than the supervisor data.

3.3 Discussion

The tool PDFHighlighter can directly label keywords in the publication, which saves lots of manual effort. However, it can be found during labeling that the tool performed best on the well-formatted PDF file. In the case of old publications with a cluttered typography or not in the form of PDF, extra manual work is needed.

Based on the result of cluster analysis (Table 3) that takes into account of both extraction efficiency and accuracy, tool performance on the label 'Parameter', 'Result for the parameter', 'Type of Test System' was always comparatively worse. The value of the label 'CONTROLS

REFERENCE ITEMS' is also low. However, the prediction score alone indicates that the extraction of the control substance is of high accuracy. The lower value in this case is due to the low appearance of the names of control substances in the publications. The extraction result for the other labels provides essential information for an AOP mapping, however without the support of the information from 'Parameter' and 'Result for the parameter', it would be difficult to confirm the relationship between up-stream and down-stream KE. Therefore, the tool SpaCy alone cannot fully meet the need for AOP mapping. Besides, the model was retrained and evaluated totally on the dioxins case, it is doubtful whether the trained model can be applied to other cases.

CHAPTER 4. DATA VALIDATION

4.1 Materials and methods

4.1.1 Platform introduction

In the stage of data validation, all the tests were performed on the OpenAI playground. The playground is powered by the GPT-3 models, which is the most recent language model coming from the OpenAI research lab team. These models include Davinci, Curie, Babbage and Ada, the rankings are reduced according to their capacity. In the following sections, the ‘GPT-3 model’ only refers to the Davinci-003 model for text training[32].

The establishment of the GPT-3 models is based on the previous study on a model called ‘Transformer’, which is a method to transform the input sequence into the output sequence following specific rules. What makes the GPT-3 models unique is that they consist of 175 billion parameters, much larger than the previously released NLP models such as the GPT-2 model and T-NLG model[33]. Because of this, the GPT-3 model is able to perform the ‘few-shot’, ‘one-shot’ or even ‘zero-shot’ tasks, i.e. complex tasks with few/one/no examples to control the form and content of the output.

4.1.2 Word embedding and data validation

Word embedding is an algorithm to encode and decode the natural language so that words with similar meanings can be grouped. Based on the introduction by OpenAI documentation[34], the word embedding is most commonly used for: search (where results are ranked by relevance to a query string); clustering (where text strings are grouped by similarity); recommendations (where items with related text strings are recommended); anomaly detection (where outliers with little relatedness are identified); diversity measurement (where similarity distributions are analyzed); classification (where text strings are classified by their most similar label).

The data validation can be seen as relating the expression in a reliability-related question with similar words from the publications. This process is mainly involving the clustering, searching and recommendation functions but other functions would also be tested when the model is generating an answer. Therefore, the use of word embedding can be a good solution for data validation.

4.1.3 Parameter setting

The temperature ranging from 0 to 1 determines the degree of randomness in the output. A temperature value of 0 results in repetitive output however sometimes leads to short and inaccurate outputs. The temperature value in this thesis was thereby set as 0.7. The maximum length of the output, to make the answer complete, was set at 300 tokens.

Other parameters including top P (the threshold for the likelihood of prediction), frequency penalty (tendency to repeat predictions), presence penalty (tendency to make new predictions) all followed the default settings.

4.1.4 Question generation methods

As discussed in the previous section, the use of word embedding can be a good solution for data validation. The approach is to use the word embedding model to generate answers to the questions, and then compare those answers to known correct answers to see how well the model is able to provide accurate responses.

During the generation of appropriate questions for the word embedding test, some criteria were followed, which include:

A) The questions should be well-formed and grammatically correct. This will help ensure that the questions are easy to understand and answer.

B) The questions should cover a wide range of topics and use a variety of different words and phrases. This will help ensure that the word embedding model is able to generalize to a wide range of contexts[35].

C) The difficulty of the questions should be appropriate. If the questions are too easy, the model may not be challenged enough to learn effectively. On the other hand, if the questions are too difficult, the model may not be able to understand them at all.

D) The questions should be able to test the quality of toxicological publications from the following aspects: bias due to deviations from intended interventions; bias due to missing outcome data; bias in the measurement of the outcome; bias in the selection of the reported result[36].

E) The questions should be designed to test a wide range of these toxicological aspects, testing the model’s ability to capture subtle differences in meaning and context[37].

In order to test whether the model can tell the difference between similar words, different levels of questions are designed for some specific aspect of the publication reliability, so that the corresponding machine-generated responses can be compared with each other.

4.1.5 Data collection

Due to the limitation of tokens, only the ‘materials and methods’ and the ‘result’ parts were imported into the model and tested. After inputting these two parts into the model, the questions from the Appendix B (Figure 1-4) can be typed into the model to start testing.

Because of the model’s deficiency of doing calculations, only qualitative data can be obtained by summarizing the quality of the output. The answers were evaluated with regard to the accuracy of the judgement, relevance to the information from the publications, completeness of the answers and the fluency of the statement.

4.2 Results

The model does not need any instructions or examples to generate answers. And the answers are automatically generated in the form of a yes/no judgement followed by a detailed explanation. In general, it is difficult to tell any difference between machine-generated answers and human language. The model can correctly identify the part of the publication an expected answer should come from, and then combine scattered information to form a sensible answer. When splitting a question into two parts, the model can still find the right corresponding words regarding different aspects from the publication (Figure 2 in Appendix B). This indicates the model can understand the relationship between words in the questions and the answers.

In most cases, the selection of words to answer the questions was mostly logical and reasonable. But in some cases where background information is not mentioned, the response of the model indicates that it is not aware of some common sense. For example, it tends to assume people who conducted the experiment are different from who handled the experimental data.

Aside from these, some other deviations between the output and the expected answer were found during testing. The deviation can come from the understanding of some abstract words of the model. For example, according to the Cambridge Dictionary, the word ‘measurement’ can be defined as both ‘the act or process of measuring’ and ‘a value, discovered by measuring, that corresponds to the size, shape, quality, etc. of something’. When asking the model ‘Does the measurement or ascertainment of outcome differ between groups?’, we want the machine to answer whether the researchers used the same method when obtaining the outcome. In other words, the word ‘measurement’ here is indicating ‘the action of measuring something’. But based on the machine-generated answer, the word ‘measurement’ was only defined as ‘the evaluation of the data obtained from the experiments’ and therefore the model output is more related to the specific value of the result instead of the method. This may indicate the model is prone to misunderstand some abstract concepts and can hardly change the pre-definition in the model.

Besides, it can be found in some rare cases that the model can sometimes use less relevant content or even make up facts to answer the question. The lowly relevant answer is probably due to the lack of hint in the question and can be easily solved by suggesting which aspect should the answer focus on. The fictive answers, on the other hand, is rare but would add difficulty when judging the quality of the output. For example, when asked whether the outcome assessment is reliable, the model made up of a non-existent third-party inspector in addition to the assessment process in the publication. In general, such kind of deviations happens when the model got a poor understanding of some specific word either in the publication or in the question.

4.3 Discussion

According to the experience in model testing, the model can mostly output detailed and reasonable answers with regard to scientific reliability. And the model is especially good at collecting information that does not need extra background information to understand. Although most of the GPT-3 output appeared very natural to humans, it is difficult to distinguish the low-quality/wrong answers from the high-quality ones because of the lack of signs. By comparison, the GPT-3 model works worse when there lacks extra background information or a clear definition of the abstract word. This is possibly due to the fact that the GPT-3 model was trained on databases such as Wikipedia which provide mostly factoid knowledge rather than common sense. And it is common in the unstructured text that some commonsense knowledge is omitted. For example, we seldom mention in a scientific publication that the same person conducted the experiments and processed the data. Therefore, the GPT-3 model can hardly get trained in this aspect. And it would be more suitable to apply this model in identifying and combining information that is clearly stated in the publications without latent background knowledge.

Because of the upper limit of inputted tokens and the model's deficiency in exploiting latent information, during the data validation of one publication, it is difficult to link the knowledge in one publication to other publications. However, comparing the NAMs result with the human/animal-related experiment result is an important way to build up scientific confidence in NAMs. Therefore, this drawback is likely to undermine the reliability of the NAM publications.

It should also be noted that in this part, only qualitative evaluation was performed. To further evaluate the machine-generated answers quantitatively, metrics that are testing the similarity between answers can be applied to test the model's ability to identify synonyms. The cosine similarity, for example, can be a choice because it disregards differences in sentence length and focuses on the semantic similarity[38]. However, coding to get the similarity value is out of the scope in this thesis, and the GPT-3 model turned out to perform poorly with regard to numerical operation.

CHAPTER 5. TOOL IMPROVEMENT

5.1 Feasibility of using the GPT-3 model for data extraction

The GPT-3 model is solely based on the Transformer network architecture, while the core of the Transformer model is the attention mechanism. The algorithm of the attention mechanism allows the sift through the mass of information and focus on a small amount of important information, ignoring most of the unimportant information[39]. Compared with other previous algorithms, this mechanism improves the model performance in long sentences where the keywords are often far from each other or have a complex structure[40]. According to the study of Ashish Vaswani et al.[41], the attention mechanism successfully addressed different tasks and also appear to exhibit behaviour related to the syntactic and semantic structure of the sentences, for example correctly interpret the polysemous words in a sentence.

On the other hand, the scientific publications more often consist of ‘complex noun phrases with many modifiers, specialized vocabulary and nominalizations that remove the agent’[42], making the sentences rather complex in linguistic structure. Also, it is clear from experience that terms or abbreviations used in the scientific publications may be interpreted in multiple senses[42]. Therefore, it can be expected that the attention mechanism of the GPT-3 model facilitates the identification and extraction of keywords from a toxicological publication.

5.2 Methods

5.2.1 Create the prompt

In the pre-experiment, the extraction instruction was given to the model by simply asking the model to categorize keywords from a sentence without any example. It turned out that it is rather difficult to create the desired form of answer (in a list with correct labels). This indicates that the so-called ‘zero-shot’ method is not suitable for the task.

During the creation of prompts, the following points should be considered in the manually created prompts. These points include the format and the content of the output, the possibility of debugging based on the output and the variance in the prompt that can lead to inconsistency in output.

Based on these, the prompt that regulates the output is designed as:

Q: Categorize the sentence ‘*sentence 1*’ in a toxicological way.

A: *word 1* [*label 1*]

word 2 [*label 2*] . . .

In the example prompt, the labels are inherited from the full-text review section, and the order of the extracted words is determined by the order in which they appear in the sentence. After typing the first Q&A pair into the model as the prompt, the second question can be input at the same time as the testing set. Once the model gives its answer to this question, the new Q&A pair will be added to the prompt list in the next evaluation (Figure 9).

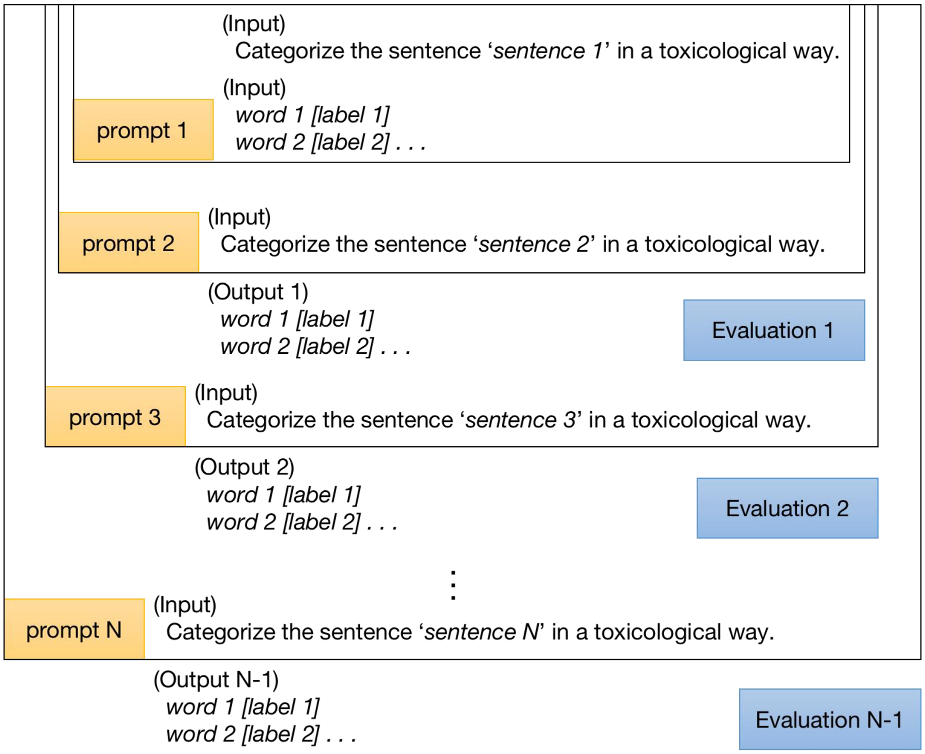


Figure 9 The workflows of prompt creation and evaluation in data extraction on GPT-3 model

5.2.2 Data Acquisition

The sentences and paragraphs for testing were acquired from the labelled publications in the stage of full-text review. In order to train the model efficiently, sentences and paragraphs with more labels were preferred.

The process of machine learning often leads to varying performance results. In order to decrease the bias from one evaluation, every testing question was asked three times. For each question to GPT-3 in an evaluation run, 3 iterations were performed to generate the answer. And the relevant information was collected after each iteration. The information acquired from the machine-generated answer should include the number of tokens in the answer, and the predicted possibility of generation on each token. By comparing the machine-generated answers with the manually labelled answers in the full-text review, the prediction scores can be calculated.

5.2.3 Data analysis

To evaluate the performance of the GPT-3 model, the three metrics, i.e. precision score, recall score and the F1 score, were also applied in this section. However, the output of the GPT-3 model as the plain text can hardly be counted individually as separate publication. Considering that the model is using the token as the minimal unit to process the input and to generate the output, the evaluation will also be performed based on tokens.

Based on the definition of the metrics, the precision score in this case is calculated as: $\frac{\sum \text{Number of tokens} \times \text{Predicted possibility}}{\text{Number of tokens in predicted answer}}$, while the recall score is calculated as: $\frac{\sum \text{Number of tokens} \times \text{Predicted possibility}}{\text{Number of tokens in right answer}}$. Both the number of tokens and the predicted possibility can be seen directly from the output of the model. The variance was also calculated on the F1 score. In order to reduce the bias from taking samples, 3 iterations of each Q&A pair were performed. For each iteration data were collected to get the mean and variance.

5.3 Results

Due to the current setting of the model, the GPT-3 only allows 4000 tokens to operate. Therefore, in total 15 rounds of the evaluation were performed and analyzed. The first 10 rounds were about extracting keywords from one specific sentence, while the last 5 were about paragraph(s) extraction.

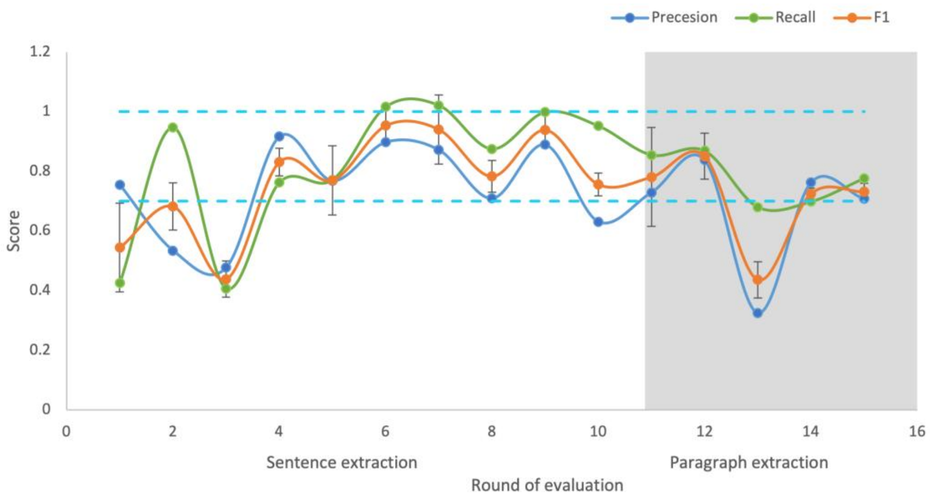
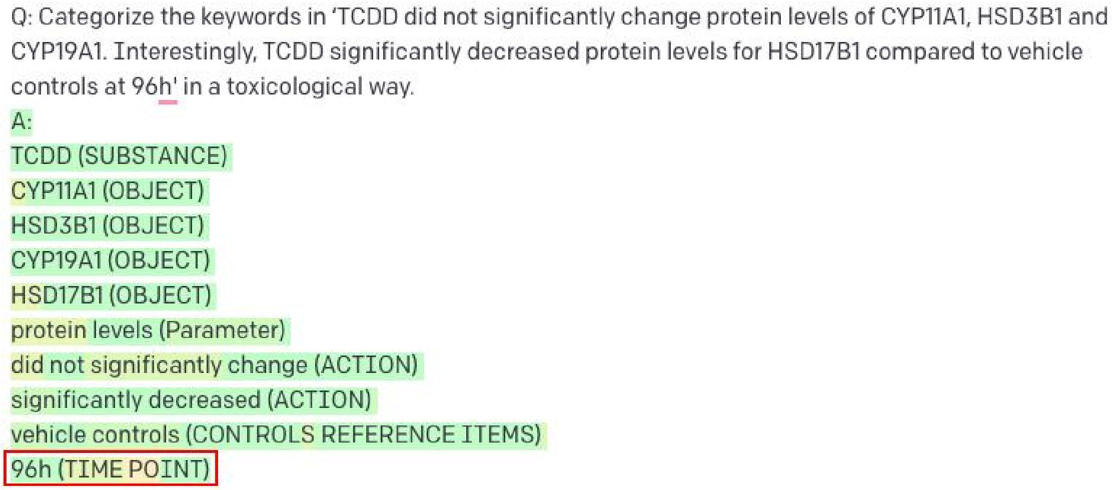


Figure 10 GPT-3 performance in data extraction

Without any previous training, the first three trials, with respectively 1-3 prompts, showed an imbalanced performance with regard to recall and precision score. At first, the precision score was far above the recall score. In the next two trials, the recall score first increased at the expense of precision score, then the model adjusted to make the two metrics closer. The instability in prediction scores also appeared in the first 3 training in the paragraph extraction. After the first three training, the three metrics were adjusted to stay within the acceptable range of 0.7-1, and the F1 score was generally more stable than the other two metrics. During the entire training, the recall score is more often significantly (p<0.05) higher than the precision score. It should also be noted that in both sentence and paragraph extraction stages, at the beginning of training, there was more the variance than the following training and the instability of precision scores with at least three prompts.

The calculation process of prediction scores in GPT-3 extraction is slightly different from those in the previous chapter, where the prediction scores are in the range of 0-1. However, in this case, the recall score occasionally went over 1. This is because the recall score, indicating the percentage of predicted positive data among all the positive data, is calculated on two different groups of tokens. The manually labelled text which provides the number of tokens in right answer is sometimes fewer than the predicted tokens. The extra tokens from the predicted content mainly come from two sources: the wrong labelled content and the newly generated labels.

The generation of new labels is one of the interesting features of the GPT-3 model. It can be found occasionally that the model output some labels that never appeared in the prompts. For example, the model would label the exposure time (e.g. 24h, 48h and 96h) as the TIME POINT (Figure 11) or label the testing techniques (e.g. RT-qPCR, microarrays) as the METHOD.



Q: Categorize the keywords in 'TCDD did not significantly change protein levels of CYP11A1, HSD3B1 and CYP19A1. Interestingly, TCDD significantly decreased protein levels for HSD17B1 compared to vehicle controls at 96h' in a toxicological way.

A:
TCDD (SUBSTANCE)
CYP11A1 (OBJECT)
HSD3B1 (OBJECT)
CYP19A1 (OBJECT)
HSD17B1 (OBJECT)
protein levels (Parameter)
did not significantly change (ACTION)
significantly decreased (ACTION)
vehicle controls (CONTROLS REFERENCE ITEMS)
96h (TIME POINT)

Figure 11 Example output with an undefined label generated

Besides, it can also be seen during the training that the model has a generally good performance on labels where the SpaCy had poor performance (PARAMETER, RESULT FOR PARAMETER). But sometimes the model combined two other labels (e.g. ALP (OBJECT) and activity (PROCESS)) into the label PARAMETER. This may be due to the current imperfection in recognizing semantic differences and the inconsistency within the manually labelled entities as the training input.

5.4 Discussion

In general, the result of the GPT-3 extraction is promising regarding the form and content of the output.

First, the input of prompts gives regulations to both the content and the form of the model output. With prompts in the desired form, the output can be automatically formed into a list with keywords and corresponding labels as is shown in Figure 11. Compared with the two-step extraction and sorting in the full-text review, it saves time and manual works. Second, the model neglected the repetition of the same word in the testing question automatically, while in the full-text review stage, every related word was labelled and extracted. This omission, however, is unlikely to result in ambiguity. This is because the order of the words in the answer suggests the sentence from which the word was extracted. Besides, different words that have the same meaning (e.g. suppress and suppression) only appeared once in one answer. This is a sign that the tool has a certain level of semantic understanding. Third, considering the number of prompts needed to train the model to complete a new task, the GPT-3 model showed a rapid speed of learning in the few-shot method. Based on the testing data, the GPT-3 model can save a large amount of manual work in labelling. Aside from these, the trained model was also tested on other chemicals, and the accurate performance shows that the model has some migration ability, or in other word, the model is not over-fitting so that it can be applied to different case studies.

Limitation and expectation

Although the GPT-3 model has the potential for data extraction, there still exist some limitations that prevent its application.

The first obstacle is the limited number of tokens which prevents the model to extract data from an entire publication. Because of technology and safety reasons, the GPT-3 model only allows 4000 tokens for input and output. This number, however, should include the tokens for model training and applying, which makes the extraction from the whole publication difficult to realize. This drawback undermines the model's ability to be used in practice. It would be more suitable for practical use if the model can be retrained for specially the full-text review task based on the GPT-3 model in advance. This method is the so-called ‘model fine-tuning’. It turned out that the fine-tuned model can not only be adjusted to perform the ‘zero-shot’ task, which greatly save the number of tokens, but also predict with higher accuracy[43].

The second limitation is the inevitable variance in the output. As is said, the temperature of the model is set as the default value of 0.7, adding a certain level of randomness in the output. While the temperature of 0 can result in repetitive but short and inaccurate outputs, the value between 0 and 0.7 would be a choice to still have some creativity but not the repetitive answers. And according to the result, the variance in the output gradually decreased as the tool performance improved. Therefore, it should be assumed that the output of a fine-tuned model is generally stable with some slight difference each time.

It should be noted that some advantages of the GPT-3 model come with potential disadvantages in future practical use. For example, the model outputs labels and the corresponding words in the form of a list at the same time. On one hand, it saves the work of further sorting and organizing. On the other hand, it adds difficulty to checking and improving the accuracy of the output. The result suggests that if the model is to be applied in real-life cases, the manually labelled data for model training should be of high quality. Manual work may be devoted to reducing the contradictions in training data and conducting an essential harmonization step afterwards. A recent study[44] also suggests that this pitfall can be (partially) resolved by degrading the question into step-by-step small questions that can reveal the answer generation process. It is possible to perform this step for labels in a publication that show a worse extraction performance during iterations. Compared with other methods, this method can be effective but time-consuming. The generation of new labels is also a two-sided feature. In the current step, the machine-designed labels are seen as unnecessary for a primitive AOP mapping. However, considering their high accuracy and the high relevance to quantitative analysis, the new labels can be further used as evidence for the KER validation.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

1. The result from the initial review indicated that the performance of Sysrev improved over iterations and that the model update resulted in a noticeable but short performance improvement. This phenomenon reflected the fact that the result of a project is also influenced by the other ongoing projects on the same platform.
2. It can be seen from the full-text review result that different extraction methods led to a significant difference in the prediction scores, but different retraining methods showed no significant difference. During the extraction, some labels (‘Action’, ‘Object’, ‘Substance’, ‘Process’, ‘Test system identity’) tend to be extracted more efficiently, while information about parameters can hardly be extracted.
3. Based on the result from data validation, the GPT-3 model showed the capacity to recognize the semantic difference in questions and provide detailed corresponding answers.
4. The GPT-3 model also has the potential for data extraction with appropriate prompts. It showed a rapid learning speed, and the output became stable and accurate over iterations. Compared with the tool performance in the full-text review, the GPT-3 model had a generally better performance in extracting parameter-related entities.

6.2 Recommendations

1. In order to improve the performance of Sysrev in the initial review, one possible way is to increase the quality of data over the platform. Another method would be based on the data from the specific project. By applying an optimized searching method to the critical labels, the dataset can be more balanced so that the classification task can be better performed.
2. The GPT-3 model can be a good complement to the SpaCy model in full-text review, especially for labels where the model showed a comparably lower extraction performance. The output of undefined labels also had the potential to be further included and processed for a quantitative AOP mapping. However, to put

GPT-3 model into practical use, it should first overcome the tokens limitation and reduce the contradictions in training data.

APPENDIX A LIST OF OPTIMIZED SEARCH TERMS

Table 1 List of optimized search terms used to direct the search to publications containing NAM-derived data

Method-related search terms
in vitro
in chemico
in silico
QSAR
zebrafish
danio rerio
c. elegans
transcriptom
cell culture
organ chip
ex vivo
cell line
organoid
drosophila melanogaster

APPENDIX B FIGURES OF QUESTIONS DESIGNED FOR TESTING THE WORD EMBEDDING FUNCTION

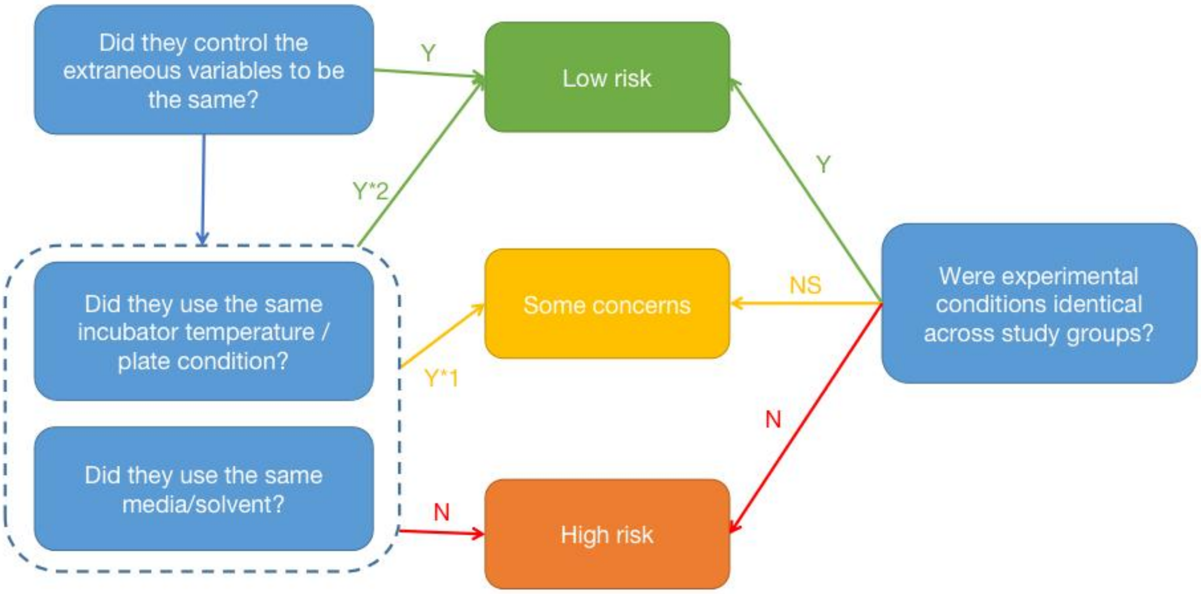

Did they control the extraneous variables to be the same?
Y
Low risk
Y*2
Y
Did they use the same incubator temperature / plate condition?
Some concerns
NS
Were experimental conditions identical across study groups?
Y*1
N
Did they use the same media/solvent?
N
High risk

Figure 1 Questions for word embedding test about experimental condition

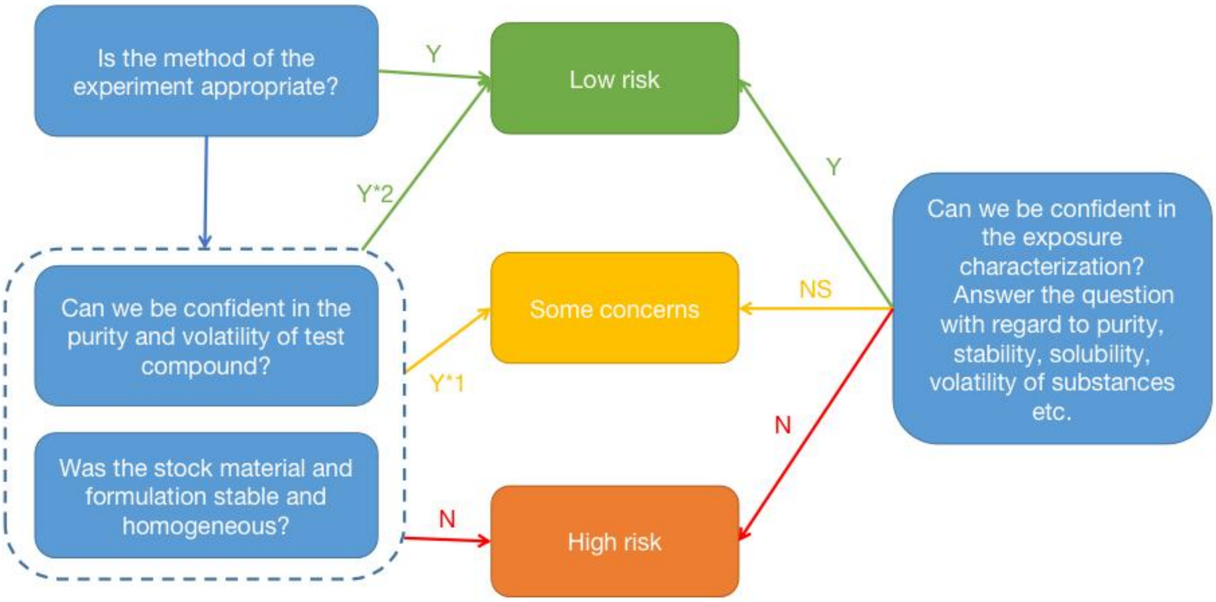

Is the method of the experiment appropriate?
Y
Low risk
Y*2
Y
Can we be confident in the purity and volatility of test compound?
Some concerns
NS
Can we be confident in the exposure characterization? Answer the question with regard to purity, stability, solubility, volatility of substances etc.
Y*1
N
Was the stock material and formulation stable and homogeneous?
N
High risk

Figure 2 Questions for word embedding test about exposure characterization

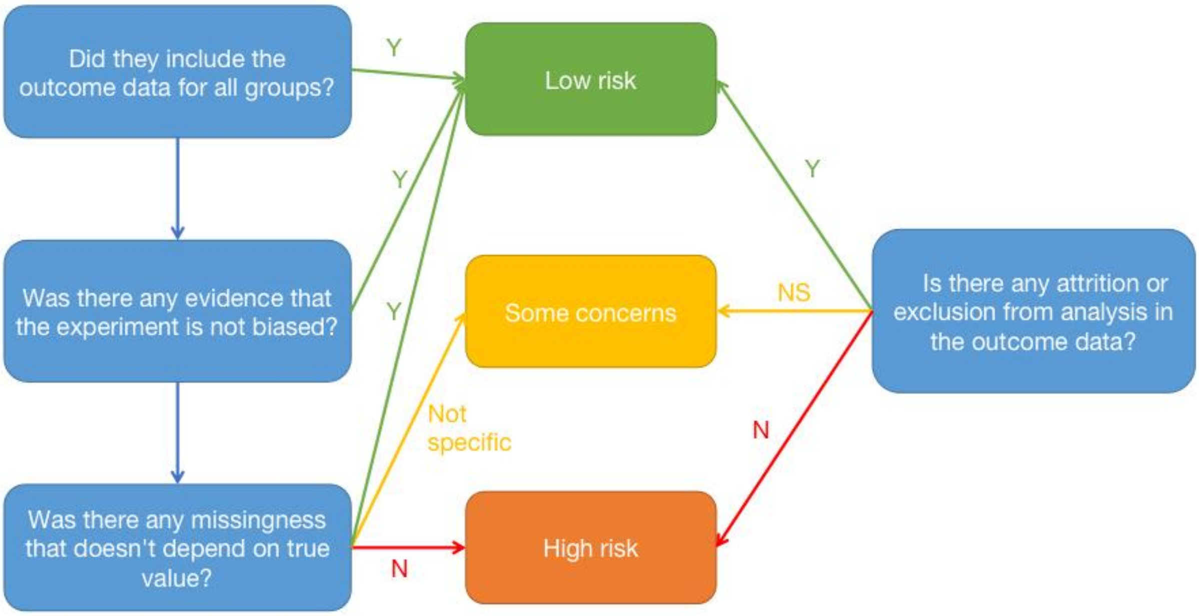

Did they include the outcome data for all groups?
Y
Low risk
Y
Was there any evidence that the experiment is not biased?
Y
Some concerns
NS
Is there any attrition or exclusion from analysis in the outcome data?
Not specific
N
Was there any missingness that doesn't depend on true value?
N
High risk
Y

Figure 3 Questions for word embedding test about outcome result

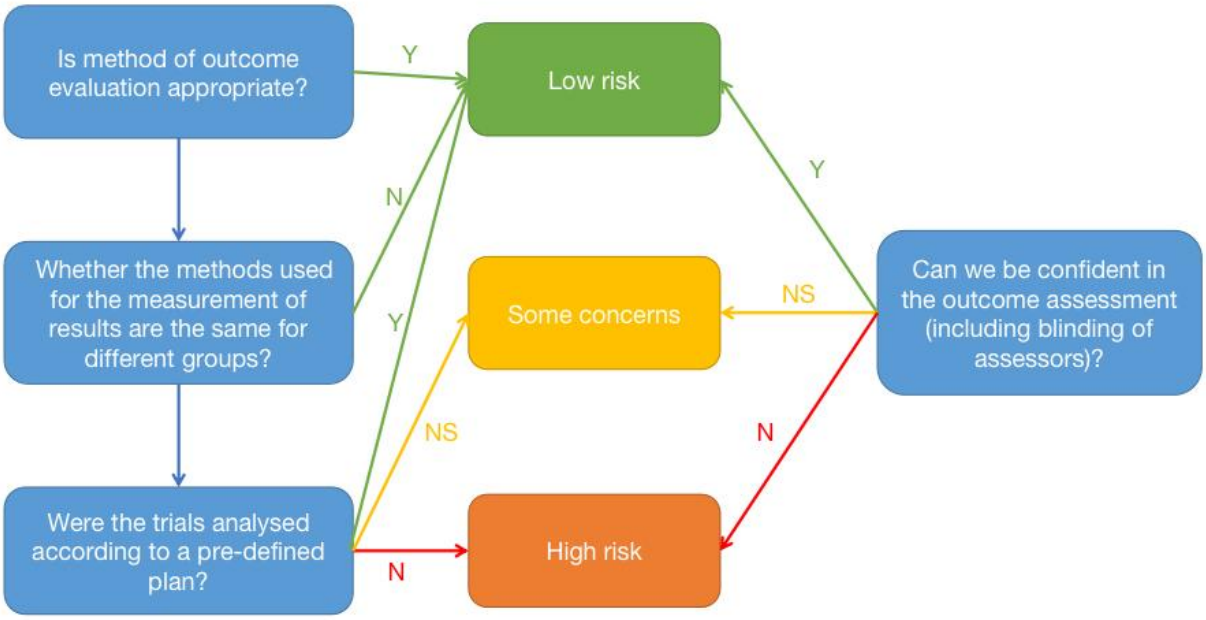

Is method of outcome evaluation appropriate?
Y
Low risk
N
Y
Whether the methods used for the measurement of results are the same for different groups?
Y
Some concerns
NS
Can we be confident in the outcome assessment (including blinding of assessors)?
NS
N
Were the trials analysed according to a pre-defined plan?
N
High risk

Figure 4 Questions for word embedding test about outcome analysis

REFERENCES

1. Escher, S., et al., *Development of a Roadmap for Action on New Approach Methodologies in Risk Assessment.* EFSA Supporting Publications, 2022. **19**.
2. Wojewodzic, M.W. and M. Andreassen, *Nordic Workshop on New Approach Methodologies (NAMs) : for Grouping and Read-Across under REACH and CLP*. TemaNord. 2022, Copenhagen: Nordisk Ministerråd. 86.
3. Patlewicz, G., et al., *Use of category approaches, read-across and (Q)SAR: General considerations.* Regulatory Toxicology and Pharmacology, 2013. **67**(1): p. 1-12.
4. Bal-Price, A. and M.E.B. Meek, *Adverse outcome pathways: Application to enhance mechanistic understanding of neurotoxicity.* Pharmacol Ther, 2017. **179**: p. 84-95.
5. Corradi, M.P.F., et al., *Natural language processing in toxicology: Delineating adverse outcome pathways and guiding the application of new approach methodologies.* Biomaterials and Biosystems, 2022. **7**: p. 100061.
6. Vinken, M., et al., *Adverse outcome pathways: a concise introduction for toxicologists.* Arch Toxicol, 2017. **91**(11): p. 3697-3707.
7. Knapen, D., et al., *Adverse outcome pathway networks I: Development and applications.* Environ Toxicol Chem, 2018. **37**(6): p. 1723-1733.
8. Spinu, N., et al., *Quantitative adverse outcome pathway (qAOP) models for toxicity prediction.* Arch Toxicol, 2020. **94**(5): p. 1497-1510.
9. Spinu, N., et al., *Development and analysis of an adverse outcome pathway network for human neurotoxicity.* Arch Toxicol, 2019. **93**(10): p. 2759-2772.
10. Kleinstreuer, N.C., et al., *Adverse outcome pathways: From research to regulation scientific workshop report.* Regulatory Toxicology and Pharmacology, 2016. **76**: p. 39-50.
11. Paini, A., et al., *Towards a qAOP framework for predictive toxicology - Linking data to decisions.* Computational Toxicology, 2022. **21**: p. 100195.
12. Bundesinstitut für Risikobewertung, d.-f.G., Wageningen University & Research, *AI4NAMS - Deliverable 1: Review of the state of the art for AI on NAMs*. 2022.
13. Braun, J.M., *Early-life exposure to EDCs: role in childhood obesity and neurodevelopment.* Nat Rev Endocrinol, 2017. **13**(3): p. 161-173.
14. Marinković, N., et al., *Dioxins and human toxicity.* Arh Hig Rada Toksikol, 2010. **61**(4): p. 445-53.
15. Wittwehr, C., et al., *Artificial Intelligence for chemical risk assessment.* Comput Toxicol, 2020. **13**: p. 100114.
16. Richarz, A.-N., *CHAPTER 1 Big Data in Predictive Toxicology: Challenges, Opportunities and Perspectives*, in *Big Data in Predictive Toxicology*. 2020, The Royal Society of Chemistry. p. 1-37.
17. Boobis, A.R., et al., *IPCS framework for analyzing the relevance of a cancer mode of action for humans.* Crit Rev Toxicol, 2006. **36**(10): p. 781-92.
18. Villeneuve, D.L., et al., *Adverse outcome pathway development II: best practices.* Toxicological Sciences, 2014. **142**(2): p. 321-330.
19. Wittwehr, C., et al., *How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology.* Toxicological Sciences, 2017. **155**(2): p. 326-336.
20. Becker, R.A., et al., *Increasing Scientific Confidence in Adverse Outcome Pathways: Application of Tailored Bradford-Hill Considerations for Evaluating Weight of Evidence.* Regul Toxicol Pharmacol, 2015. **72**(3): p. 514-37.

21. Carusi, A., et al., *Harvesting the promise of AOPs: An assessment and recommendations.* Science of the Total Environment, 2018. **628**: p. 1542-1556.
22. Leist, M., et al., *Adverse outcome pathways: opportunities, limitations and open questions.* Archives of toxicology, 2017. **91**(11): p. 3477-3505.
23. Bundesinstitut für Risikobewertung, d.-f.G., Wageningen University & Research, *AI4NAMS - Deliverable 4: Final study protocol per case study – General methodology*. 2022.
24. van der Zalm, A.J., et al., *A framework for establishing scientific confidence in new approach methodologies.* Archives of Toxicology, 2022. **96**(11): p. 2865-2879.
25. Bozada, T., Jr., et al., *Sysrev: A FAIR Platform for Data Curation and Systematic Evidence Review.* Front Artif Intell, 2021. **4**: p. 685298.
26. Bundesinstitut für Risikobewertung, d.-f.G., Wageningen University & Research, *AI4NAMS - Deliverable 5: Draft case study report Case study 2 - CC5 - Phthalates*. 2022.
27. Bundesinstitut für Risikobewertung, d.-f.G., Wageningen University & Research, *AI4NAMS - Deliverable 2: List of possible case studies.* 2022.
28. Rohatgi, A., *WebPlotDigitizer user manual version 3.4.* URL http://arohatgi. info/WebPlotDigitizer/app, 2014: p. 1-18.
29. Buckland, M. and F. Gey, *The relationship between Recall and Precision.* Journal of the American Society for Information Science, 1994. **45**(1): p. 12-19.
30. Cohen, J., *Statistical power analysis for the behavioral sciences*. 2013: Routledge.
31. Sun, Y., A.K. Wong, and M.S. Kamel, *Classification of imbalanced data: A review.* International journal of pattern recognition and artificial intelligence, 2009. **23**(04): p. 687-719.
32. Bundesinstitut für Risikobewertung, d.-f.G., Wageningen University & Research, *AI4NAMS - Deliverable 5: Draft case study report - Case study 3 – EC1 - Hypothyroidism.* 2022.
33. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
34. *OpenAI API.* [cited 2023 January 21]; Available from: https://beta.openai.com/docs/guides/embeddings/limitations-risks
35. Hollenstein, N., et al., *CogniVal: A framework for cognitive word embedding evaluation.* arXiv preprint arXiv:1909.09001, 2019.
36. Higgins, J.P., et al., *Assessing risk of bias in a randomized trial*, in *Cochrane Handbook for Systematic Reviews of Interventions*. 2019. p. 205-228.
37. Wang, B., et al., *Evaluating word embedding models: methods and experimental results.* APSIPA transactions on signal and information processing, 2019. **8**.
38. Faruqui, M., et al., *Problems with evaluation of word embeddings using word similarity tasks.* arXiv preprint arXiv:1605.02276, 2016.
39. Manning, C.D., et al., *Emergent linguistic structure in artificial neural networks trained by self-supervision.* Proceedings of the National Academy of Sciences, 2020. **117**(48): p. 30046-30054.
40. Vig, J. and Y. Belinkov. *Analyzing the Structure of Attention in a Transformer Language Model*. 2019. Florence, Italy: Association for Computational Linguistics.
41. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
42. Persson, T., Å. af Geijerstam, and C. Liberg, *Features and Functions of Scientific Language(s) in TIMSS 2011.* NorDiNa: Nordic Studies in Science Education, 2016. **12**(2): p. 176-196.

43. Cobbe, K., et al., *Training verifiers to solve math word problems.* arXiv preprint arXiv:2110.14168, 2021.
44. Reppert, J., et al., *Iterated Decomposition: Improving Science Q&A by Supervising Reasoning Processes.* arXiv preprint arXiv:2301.01751, 2023.