Finding unknown stimulants applied in food supplements using Artificial Intelligence

Anand Gavai (✉ anand.gavai@wur.nl)  Wageningen Food Safety Research (WFSR)  https://orcid.org/0000-0002-4738-190X

Yamine Bouzembrak  Wageningen Food Safety Research (WFSR)

Leonieke Bulk  Wageningen Food Safety Research (WFSR)

Ningjing Liu  Wageningen Food Safety Research (WFSR)

Lennert van Overbeeke  Wageningen Food Safety Research (WFSR)

Lucas Heuvel  Wageningen Food Safety Research (WFSR)

Hans Mol  Wageningen Food Safety Research (WFSR)

Hans Marvin  Wageningen University and Research (WUR)

Article

Keywords: Stimulants, Enhancers, Social media, MedISys, Word Embedding, Text Mining, Emerging Risk.

DOI: https://doi.org/10.21203/rs.3.rs-447092/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

The world market for food supplements is large and is driven by the claims of these products to, for example, treat obesity, increase focus and alertness, decrease appetite, decrease the need for sleep or reduce impulsivity. The use of illegal compounds in food supplements is a continuous threat, certainly because these compounds and products have not been tested for safety by competent authorities. It is therefore of the utmost importance for the competent authorities to know when new products are being marketed and to warn users against potential health risks.

In this study we present an approach to detect new and unknown stimulants in food supplements using machine learning. More than 20 new stimulants were identified from two different data sources, namely scientific literature applying word embedding on >2 million abstracts and articles from formal and social media on the world wide web using text mining.

The results show that the developed approach may be suitable to detect “unknowns” in the emerging risk identification activities performed by the competent authorities, which is currently a major hurdle.

1. Introduction

The global dietary supplements market size was estimated at USD 140.3 billion in 2020 and is expected to expand at an annual growth rate of 8.6% from 2021 to 2028. Factors, such as rising health concerns and the changing lifestyles and dietary habits have been driving this growth in demand. Consumers find supplements attractive to compensate for imbalances of nutrients in their diet or unhealthy lifestyle, and to prevent chronic diseases, among others (Biesterbos, et al., 2019). Claims about the benefits of food supplements, and the marketing thereof, are regulated in Europe through directives such as (Ref 2002/46/EC). Food supplements include products such as vitamins, energy drinks, protein drinks, weight loss supplements and exotic or novel foods. A subgroup of food supplements are stimulants, which are agents (e.g., drugs) that produce a temporary increase of the functional activity or efficiency of an organism. Often in the consumer market they are used to treat obesity, increase focus and alertness, decrease appetite, decrease need for sleep (Carroll, et al., 2006). Although these compounds are legally regulated, illegal compounds are also often sold as food stimulants. Another example concerns undeclared anabolic steroids in sports supplements such as the banned substance 1,3-DMAA being marketed as an extract of Aconitum kusnezoffii (Cohen, et al., 2018). While their consumption may have the intended effect of increasing the muscle mass for an unaware user, serious adverse effects are common (Martin, et al., 2018). Research in this field also validates the use of online forums as a source of new leads to illegal practices (Helle, et al., 2019). Not only well-known enhancers are illegally added to supplements, but experimental or even prohibited substances may be used (Cohen, et al., 2018). Because of its market potential and difficulty to control, an increase in adulteration (adding synthetic compounds or illicit herbal materials) has been observed (Končić, 2018) and a further increase is expected.
To obtain an overview of adulteration of food supplements on the Dutch market, the Netherlands Food and Consumer Product Safety Authority (NVWA) analysed samples collected from 2013–2018 and observed that 64% of the samples contained one or more unauthorized pharmacological active compounds or plant toxins (Biesterbos, et al., 2019). This result demonstrates that regular monitoring of market samples is important to protect public health, but the wealth of potential compounds that can be used and the criminal aspects related to these illegal practices, makes this a growing challenge. The database used for screening the samples in the study of Biesterbos et al contained >1500 compounds (i.e. pharmaceutical substances, adulterants and plant toxins) and is continuously being expanded based on new information and reported adulterations (Biesterbos, et al., 2019).

In this study, we present a novel approach to find new compounds that can be used illegally in food supplements and which should be added to the database used for the screening. The focus was on the subcategory “stimulants” of which 428 compounds were present in the reference database.

The first data source explored was scientific literature, where the focus was on compounds that can be used in supplements and have been described in literature. For an expert, it would be unfeasible to read the overwhelming amount of scientific literature available in this topic to find new stimulants that should be added to the monitoring list. However, machine learning has made it possible to gather information automatically from text through natural language processing (NLP) techniques (Chowdhary, 2020). A word embedding model was developed to find unknown stimulants automatically from the scientific literature. A word embedding model captures words in high-dimensional vectors, called embeddings, while preserving syntactic and semantic relationships to other words (Bengio, et al., 2003; Mikolov, Corrado, et al., 2013; Pennington, et al., 2014). This results in a model in which related words are closer together in vector space. It is trained in an unsupervised way, meaning that a labeled dataset is not required. The embeddings are learned by looking at what words appear in the same context or co-occur together often. A very good example of how a word embedding model works can be found in the famous example of the embeddings of “King” - “Man” + “Woman” which results in the embedding for “Queen” (Mikolov, Yih, et al., 2013), showing that semantic information is captured by the model in a systematic way. Using such a word embedding model we can find words that co-occur together with the word “stimulant”, which will be the case for compounds that are described as stimulants in the scientific literature.

The second data source which is aimed to find new compounds that are already on the market, and of which its usage is described on the internet, is the European Media Monitor (EMM). EMM is a news aggregation service operated by the European Commission which is based on text mining, searching the world wide web (official websites, blogs etc.) for news reports 24/7 in 60 languages (Bouzembrak, et al., 2018). It consists of 3 platforms being NewsExplorer, NewsBrief, and MedISys, of which the latter displays articles with interest to public health (e.g., diseases, plant pests, psychoactive substances). In this study, MedISys was used to collect publications on new stimulants used or discussed somewhere in the world.
The approach developed in this study yielded new stimulants that potentially can or are illegally used as stimulants in food supplements and which can pose a health risk for the user. The approach was developed for stimulants in food supplements, but the methodology may be applied to any other topic. In emerging risk identification (ERI) as employed by authorities to identify food safety risks at an early stage (Marvin, et al., 2009; Meijer, et al., 2020), this approach may be suitable to be wider implemented to find “unknowns”, which is the major hurdle in ERI.

2. Materials And Methods

In this study the list of stimulants present in the reference list of Wageningen Food Safety Research (WFSR), which is used to screen samples from the Dutch Food Safety authority (NVWA), was taken as a starting point. This list was developed over a number of years and consists of 428 different compounds varying from prescription medicine to prohibited recreational drugs. “Unknown” stimulants are defined as those stimulants that are not included in this reference list.

The approach developed for the identification of unknown stimulant compounds in food supplements consisted of i) “word embedding” of the relevant scientific literature complemented with ii) text mining the world wide web using the MedISys infrastructure.

2.1. Word embedding to detect unknown stimulants from scientific literature

Data collection

The list of 428 stimulants present in the reference database, complemented with their synonyms as found in PubChem was used to collect scientific publications from Europe PMC for the period 1990–2019. Europe PMC was used as a data source because it is an open-access literature database containing over 38 million abstracts from specifically biomedical and life sciences research articles. Titles and abstracts that contained one or more of the search terms were collected, yielding a total of 2.1 million scientific articles.

Word embedding model

The word embedding model used in this study is the Word2Vec neural network variation created by Tshitoyan et al. (Tshitoyan, et al., 2019). They used the word embedding model to predict new thermoelectric materials automatically from abstracts of scientific literature. A Word2Vec model contains three layers (an input, hidden and output layer) and is trained by predicting the probability for each word in the vocabulary that it appears in the context of a specific target word. After training, the word embeddings are set to the learned weights of the hidden layer, where the word embedding of the i’th word in the vocabulary corresponds to the i’th row of the weights. The weights of the output layer are called the output embeddings, where the i’th column embeds the context words of the i’th word in the vocabulary. The code created by Tshitoyan et al. to build and train the Word2Vec model is openly available and was written using Python 3.6. Their code was used to train our own word embedding model.
The 2.1 million titles and abstracts were used as training data for the word embedding model in order to find related stimulants in the scientific literature that were not present in the list of 428 stimulants. Each title and its respective abstract were concatenated as one data point. These texts were pre-processed by removing uninformative words, like the copyright information or section information (e.g., words like introduction, conclusion) to only retain the words containing the information on the actual research. More pre-processing was done in the framework by Tsitoyan et al. in which words were deaccented and lowercased, unless the word was a chemical formula or abbreviation, and all numbers were converted to a special number token. The model was trained with the hyperparameters as set in the available framework. This meant training a skip-gram neural network with a hidden layer of size 200 with a negative sampling loss, using 15 negative samples, for 30 epochs. Training was done with an initial learning rate of 0.01 which decreased to 0.0001 over time, a context window of 8 and subsampling with a 0.0001 threshold. The Word2vec phrases were created with a phrase count of 10, a score threshold of 15 and a phrase depth of 2. From the trained word embedding model, we collected the words of which the output embeddings were closest to the word embedding of the word “stimulant” in the learned vector space to predict the words that co-occur with “stimulant” the most. As in the research by Tsitoyan et al., only words that occur more than three times in the training data were considered in order to have a more accurate representation of the data. Furthermore, since we were only interested in finding stimulants, all words that were not chemical compounds were removed from the collected set by checking the words against the PubChem database. As a last step, the stimulants from the existing reference database together with their synonyms were also removed from the set, leaving only possible new stimulants. The top 50 compounds from this set were evaluated by an expert for their validity.

2.2. A MedISys text mining model to detect unknown stimulants on the world wide web

The MedISys infrastructure does not collect publications on the world wide web on stimulants in food supplements specifically and therefore must be trained for this purpose. This includes the development of a dedicated filter to find publications of interest followed by a validation step to reduce the extent of “noise” (irrelevant publications). WFSR has special permission from the owner (Joint Research Centre) to develop filters on the MedISys infrastructure.

*Developing a filter for stimulants in food supplements on the MedISys infrastructure*

The construction of a filter in MedISys for stimulants was done according to the steps defined in (Bouzembrak, et al., 2018) and consists of the following 3 steps: (i) development of a set of keywords, (ii) creating a new filter in MedISys for stimulants based on the defined set of keywords, and (iii) evaluation and improvement of the performance of the newly developed filter.

Step 1: All stimulants present in the WFSR database (i.e. 428 chemical names) were combined with and without the words “similar”, “replace”, “stimulants”, “supplements”, “food supplements”, and “new stimulants”. Only English terms were used.

Step 2. A new filter was created in MedISys in which the developed set of keywords were integrated (Bouzembrak, et al., 2018).
Step 3. The filter was tested over a period of 6 months in which the performance (i.e. % of relevant articles collected) was examined by an expert and keywords were adjusted to improve its performance. After two iterations, the filter reached a relevance level of 78 %, which is considered optimal. This optimized filter was run for one year (July 2018 to end of June 2019) and the collected publications (i.e. 806 articles) were evaluated by an expert. Stimulants other than those in the reference set were recorded.

Since the reports collected and presented on the MedISys system are only visible as long as these reports are available on the original location, an automatic retrieval system was created that retrieves the collected reports from the MedISys website and stores them on a data infrastructure at WFSR for further analysis. To this end, a script was developed in Python 3.6, which is run on the WFSR infrastructure inside a Docker container on an OpenShift cluster. This system collects new articles from the MedISys website once every 6 hours. The data is stored in an Elasticsearch 7.1 database hosted on a cloud infrastructure at WFSR and visualised in a dashboard using Kibana software.

The data available for each report on MedISys includes the country of origin, the date, the time of collection by MedISys, the keywords present in the article, the source of the article, a link to the original website, and an automatically generated summary. If the original article was not written in English, a translation of the article title and of the automatically generated summary was produced using the Google Translate API, and stored alongside the original text.

3. Results And Discussion

3.1. Detection of unknown stimulants from the scientific literature using word embedding

From the trained word embedding model, the collection of words that in vector space were closest to the vector of the word “stimulant” were collected. In Fig. 1, a three-dimensional representation of the trained word embedding space is shown. The projection presented centralizes the word embedding for “stimulant” and shows the top 1000 nearest neighbors in color. Examples of the neighboring words are plotted next to their corresponding points in space. In Fig. 2, a two-dimensional projection of the word embeddings for “stimulant” and its closest 50 neighbors is presented. Analysing these figures reveals that the word embedding model has been successful in learning which words appear in a similar context to the word “stimulant”. From Fig. 2 it can also be seen more clearly that semantically related words get placed closer together in vector space, resulting in small clusters of similar words.

From the collections of words closest to the word “stimulant”, only the words that were chemical compounds were saved. The known stimulants from the reference database were removed and the remaining 50 highest ranking compounds (see supplement 1) were analysed for possible new stimulants by an expert in food supplement safety. Many of the top 50 compounds were upon inspection discarded as a possible new stimulant for several reasons. The first being that some of the found compounds were not meant as a compound in this context. Examples of this are “Hallucinogen” or “CNS”, both have an entry in PubChem as a synonym of a compound, but it is obvious that in this case the words have a different meaning. Hallucinogens are next to stimulants a different class of drugs, while CNS stands for...
central nervous system. For both it is very logical that the words co-occur together with the word stimulant. Other reasons for exclusion are that the compounds are not registered synonyms in PubChem for stimulants already in the known list of stimulants (e.g., Lisdexamfetamin or Cath), that they are salts or known metabolites of the existing stimulant list (e.g., N-Cyanomethylmethamphetamine or Dl-Threo-Methylphenidate) or that the compounds are not considered stimulants (Fursultiamine or 2,2,2-Trichloroethyl Chloroformate). For the latter case, the found compounds are often used together with stimulants or their structures are similar to a specific stimulant which makes them a suitable treatment for addiction of that stimulant, which explains their co-occurrence with the word stimulant. After removing the excluded compounds, a list of fourteen new stimulants is left. Of this list two stimulants were merged with other stimulants in the list, because they were synonyms of each other (2-Benzhydrylpiperidine = Desoxypipradrol and 5-(2-Aminopropyl)Indole = 5-IT). The remaining twelve stimulants were judged for the possibility of being added to food supplements. Two of the stimulants, 6-Hydroxytrypargine and Oxolinic Acid, were considered not relevant for food supplements as the former is a spider toxin and the latter is in use as an antibiotic. The other ten stimulants were considered relevant to include in the stimulant database and are shown in Table 1, including a short description
Table 1
List of newly identified unknown stimulants from scientific literature.

| Stimulant name                        | Description                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2-Benzhydrylpiperidine                | 2-Benzhydrylpiperidine, also known as desoxypipradrol or 2-DPMP, is a drug which acts as a norepinephrine-dopamine reuptake inhibitor. It is used as a recreational drug, but because of its toxicity and adverse health-effects it is already being controlled in some countries (J. M. Corkery, et al., 2012). |
| RTI-98                                | RTI-98, also known as nor-beta-CIT, is a drug closely related to cocaine. It acts as an uptake inhibitor of dopamine, norepinephrine and serotonin. RTI-98 is mainly used in scientific research as it can be used to assess the density of serotonin transporters in the brain well (Joensuu, et al., 2007; Tolliver & Carney, 1995). |
| N-methyl-2-Al                         | N-methyl-2-Al, also known as N-methyl-2-aminoisobutane or NM-2-Al, is an analogue of amphetamine, and works as a dopamine and norepinephrine releasing agent. It is being sold as a designer drug, but little is known about its toxicity (Manier, et al., 2020; Mestria, et al., 2020). |
| 5-(2-Aminopropyl)indole               | 5-(2-Aminopropyl)indole, also known as 5-IT or 5-API, is a designer drug working as a dopamine, norepinephrine and serotonin releasing agent. The compound is an indole derivative and isomer of alpha-methyltryptamine. Because of the high risk for abuse and possible adverse health effects it has been banned in most western countries (Katselou, et al., 2015; Marusich, et al., 2016). |
| Ethylphenidate                        | Ethylphenidate is a psychoactive substance that is an analogue of methylphenidate (Ritalin). It works similarly to methylphenidate and is a dopamine and norepinephrine releasing agent. It is used as a recreational stimulant. Because of severe adverse health effects, it has been banned in several countries (Maskell, et al., 2016; Parks, et al., 2015). |
| D2PM                                  | D2PM, also known as diphenylprolinol, is a psychoactive designer drug that is a norepinephrine-dopamine reuptake inhibitor. D2PM is a pyrrolidine analogue and acts similarly to cocaine. It has been established that it produces toxic effects in humans, but is still available as a 'legal high' (Wood & Dargan, 2012). |
| (+)-UH232                             | (+)-UH232 is an aminotetralin derivative. It is considered a weak stimulant and acts as a mixed agonist-antagonist for dopamine receptors. (+)-UH232 has been mainly used in scientific research as it can be used to assess the role of dopamine receptors in the brain well (Kling-Petersen, et al., 1994). |
| 7-(beta-Chloroethyl)theophylline       | 7-(Beta-Chloroethyl)theophyllin, also known as 7-(2-chloroethyl)theophylline or 7-CET, is a derivative of the natural compound theophylline and is an adenosine receptor antagonist. The effect is similar to caffeine, but more potent (Coffin & Spealman, 1989). |
| N-Methyl-3-phenyl-norbornan-2-amine    | N-Methyl-3-phenyl-norbornan-2-amine, also known as Camfetamine, is closely related to fencamfamine, but it has a stronger stimulating effect. The compound works as an indirect dopaminergic agonist. It is sold as a designer drug and is mostly unregulated. Little is known about the potential health risks related to its use (Cinosi, et al., 2014). |
### 3.2. Detection of unknown stimulants from formal and social media using MedISys

Within MedISys, a filter was created to collect publications worldwide on unknown stimulants in food supplements. This filter was applied in the period July 2018 to June 2019. The collected articles were transferred from the MedISys to a cloud infrastructure, where it is stored for further analysis. Information on the collected articles is shown in a dashboard with interconnected panels (Fig. 3). As shown in Panel 1 of Fig. 3, in this period, 806 articles were collected and a considerable variation was observed in the number of articles per week in the period analysed (Panel 2 of Fig. 3).

The articles originated from many countries (i.e. 49, see Panel 5 of Fig. 3), but the majority came from United States (68%) followed by United Kingdom (7%), India (2.6%), the Netherlands (2.6%) and Australia (1.6%). Most of the articles (i.e. 67%) were associate with keyword “similar” followed by keyword “MDMA” (47%), “stimulant” (30%) and “supplement” (12%) (Panel 3 of Fig. 3). To visualise the the content of the collected articles (i.e., titles and abstracts) a co-occurrence network visualisation was prepared (Fig. 4).

In the network, each circle represents a word and the size indicates the number of times it was mentioned in the title and abstracts. The words that co-occur often are located closer to each other in the network. Five groups can be distinguished, which are indicated in different colours in Fig. 4. The groups are centered around the words: i) use; ii) report, iii) week/hour, iv) supplement and v) ecstasy.

All articles collected were analysed by an expert on food supplements with the focus to find other stimulants than those included in the reference database. Articles that were associated with the keyword “similar” (i.e. 538) are of special interest because these may mention new, unknown stimulants. This evaluation yielded in total 27 possible unknown stimulants (see Supplement 2). Upon closer inspection some of the compounds were identified as synonyms of each other and were therefore merged together. Further assessment revealed that eleven of the remaining compounds found could not be classified as stimulants, but rather were drugs with different properties (e.g., hallucinogenic or dissociative). These compounds were consequently removed from the list of possible stimulants. Ultimately, this resulted in a final list of ten unknown stimulants and are shown in Table 2, including a short description.

| Stimulant name | Description |
|----------------|-------------|
| Paraxanthine   | Paraxanthine, also known as 1,7-Dimethylxanthine, is a derivative of xanthine and metabolite of caffeine, with similar stimulating properties. It acts as an antagonist for adenosine receptors (Benowitz, et al., 1995). |
Table 2
List of unknown stimulants collected from media on the world wide web by MedISys.

| Stimulant name | Description |
|----------------|-------------|
| 25C-NBOMe      | 25C-NBOMe, also known as N-(2-methoxybenzyl)-2-(4-chloro-2,5-dimethoxyphenyl)ethanamine, is a derivative of the phenethylamine 2C-C. Next to the hallucinogenic effects the NBOMe drugs are known for, 25C-NBOMe also has a stimulating effect comparable to MDMA. It is a partial agonist of 5-HT2A receptors. It has a high risk of acute toxicity, and fatalities by 25C-NBOMe have been reported. 25C-NBOMe has been a worldwide controlled substance since 2015 (Bersani, et al., 2014; Wohlfarth, et al., 2017). |
| 25I-NBOMe      | 25I-NBOMe, also known as 2-(4-Iodo-2,5-dimethoxyphenyl)-N-((2-methoxyphenyl)methyl)ethanamine, is a derivative of the phenethylamine 2C-I. Similar to 25C-NBOMe, although known for its hallucinogenic effects, 25I-NBOMe has been shown to have stimulating properties. It is a full agonist for the 5-HT2A receptor. Usage may lead to severe clinical toxicity in its users. 25I-NBOMe has been a worldwide controlled substance since 2015 (Hill, et al., 2013; Wohlfarth, et al., 2017). |
| 6-APB          | 6-APB, also known as 6-(2-aminopropyl)benzofuran, is a designer drug with both hallucinogenic and stimulant properties. It is both an uptake inhibitor and releasing agent of dopamine, norepinephrine and serotonin and acts as an agonist of 5-HT2A and 5-HT2B receptors. Because of the interaction with 5-HT2B receptors, 6-APB is cardiotoxic with long-term use, but also has the potential for acute toxicity. It is a controlled substance in several countries, but remains one of the most sold new psychoactive substances in Europe (Brandt, et al., 2020; Chan, et al., 2013; Roque Bravo, et al., 2020). |
| 5-APB          | 5-APB, also known as 5-(2-aminopropyl)benzofuran, is similarly to 6-APB a designer drug with hallucinogenic and stimulant properties. It is both an uptake inhibitor and releasing agent of dopamine, norepinephrine and serotonin and acts as an agonist of 5-HT2A and 5-HT2B receptors. Because of the interaction with 5-HT2B receptors, 5-APB is cardiotoxic with long-term use, but also has the potential for acute toxicity and seems to be more toxic than 6-APB. It is only controlled in a few countries. (Brandt, et al., 2020; Roque Bravo, et al., 2020; Welter, et al., 2015). |
| 5-MeO-DALT     | 5-MeO-DALT, also known as N,N-Diallyl-5-methoxytryptamine, is mostly used as a hallucinogenic drug, but drug users also report more energy, euphoria and arousal when taking it. Little information can be found in scientific literature about its exact stimulant mechanisms in humans. Research has shown, however, increased locomotor activity in rodents when administrating 5-MeO-DALT. It is a controlled substance in several countries (John M. Corkery, et al., 2012; Gatch, et al., 2017). |
| Dextromethorphan| Dextromethorphan, also called DXM, is a cough medicine which has been used since the 1950's. Abuse of dextromethorphan has been frequent, because of its stimulating and psychoactive properties. It acts as a serotonin reuptake inhibitor and is a NMDA receptor antagonist. Dextromethorphan has minimal adverse reactions at low doses, but when taken frequently and in higher doses can lead to severe intoxication (Logan, et al., 2009; Reissig, et al., 2012; Schwartz, et al., 2008). |
| Stimulant name | Description |
|---------------|-------------|
| 5-MeO-MiPT    | 5-MeO-MiPT, also called moxy, is a psychedelic with stimulant properties. It inhibits the re-uptake of 5-HT, dopamine and norepinephrine. The toxicity is still relatively unknown, but recent research showed evidence of acute toxicity in mice when given a high dose. 5-MeO-MiPT is still uncontrolled in large parts of the world (Altuncı et al., 2021; Repke et al., 1985). |
| 3-FPM         | 3-FPM, also known as 3-Fluorophenmetrazine, is a designer drug. It is a derivative of phenmetrazine. 3-FPM is a norepinephrine-dopamine releasing agent. The toxicity of 3-FPM has not been studied well at the moment of writing, although reports of severe adverse effects in human users have already been reported. It has been made a controlled substance in a few countries (Bäckberg et al., 2016; Fawzy et al., 2017; Mayer et al., 2018). |
| N-Ethylhexedrone | N-Ethylhexedrone, also known as Hexen, is a designer drug with stimulant properties similar to amphetamine. It is a synthetic cathinone and acts as a norepinephrine-dopamine reuptake inhibitor. There is limited data available in the scientific literature about the toxicity of N-ethylhexedrone in humans, but fatal N-ethylhexedrone intoxications have already been reported and recent research has shown toxicity in vitro. N-Ethylhexedrone is an internationally controlled substance since 2020 (de Mello-Sampayo et al., 2021; Domagalska et al., 2020; ECD, 2020; Majchrzak et al., 2018). |
| Methamnetamine | Methamnetamine, also known as MNA or PAL-1046, is an analog of methampethamine and has similar stimulant properties. It is being sold as a designer drug and acts as a releasing agent of serotonin, norepinephrine and dopamine. There is very little scientific literature available about methamnetamine and its toxicity, but it is currently being detected in drug screenings across Europe. Methamnetamine is an uncontrolled substance in most countries (Lajtai et al., 2020; Richeval et al., 2019; Rothman et al., 2012). |

### 3.3. Comparison of the methodologies to detect unknown stimulants

It is remarkable that both methodologies yielded completely different new stimulants, indicating that these methods are complementary. One could expect that compounds being developed and discussed in the scientific literature would proceed the application in products that are on the market. To verify this, the stimulants found online in media were first checked against the top 250 possible stimulants from the word embedding to see if they had been ranked lower than the top 50 that was checked by the expert, but were still relatively contextually close to the word “stimulant”. Only one of the stimulants appeared in the top 250: 3-Fluorophenmetrazine (3-FPM), which was number 123. Next, the remaining stimulants found in online media were searched in the corpus that the word embedding model was trained on. All stimulants except methamnetamine were present in the corpus, which is a logical consequence of the fact that there is very little scientific literature to find about methamnetamine. The frequency of being present in an abstract varied across the rest of the stimulants, but for almost all the occurrence ranged between 1–30 times. Dextromethorphan was an exception to this, the compound was present in 1293 abstracts. Dextromethorphan has been researched extensively as a treatment for a variety of health conditions or in the context of drug metabolism, but only around 20 abstracts discussed its psychoactive properties. Upon inspection of the scientific abstracts containing the stimulants found in media, it
became apparent that they were not described as stimulants, but rather as new psychoactive substances (NPS). NPS are a group of compounds often known as designer drugs or “legal highs” that can be categorized as cannabinoids, stimulants, depressants and hallucinogens (Shafi, et al., 2020). It appeared that, when scientific literature discusses new drugs that are being recreationally used or abused, which currently is where these stimulants occur most in literature, the distinction between the different categories of NPS is seldom made and their individual properties are not described in the abstract. Unless the corpus contains literature stating (indirectly) the stimulant properties of a compound or by comparing the compound to a known stimulant, the word embedding model will not associate the compound strongly with the word ‘stimulant’. In media, however, these designer drugs are often described together with their effects, which makes it easier to extract the stimulating ones. Media can thus best be used to identify stimulants that are new and up-and-coming among recreational drug users, and scientific literature can identify the stimulants that through research have been discovered to have those properties or stimulants that are established enough in the recreational drug world to have been the specific target or research.

3.4. Limitations of the approach

In the approach developed here, the reference list of the stimulant database was taken as a starting point, and therefore determines the outcome of the analysis. A word embedding model that is purely machine driven may also be too narrow to understand why and how markets tend to adapt which is generally influenced by parameters such as costs, availability of resources and the law and order situation of a country. Compounds found that are not on the reference list were considered as “unknown” in this study. It is clear that for other controlling organisations that have a different reference list, “unknown” compounds may be labeled differently. Another limitation is that in the scientific literature approach, only English literature was considered. English was also the dominant language in the online media dataset, but articles written in other languages, including Spanish, Chinese and Arabic, were also collected, translated and analysed. Many of the keywords, being chemical names or acronyms, were universally found across multiple languages, whereas other keyword such as “similar” require the addition of a suitable translation to the filter. An additional technical challenge in this field is the transliteration of characters between English and Arabic and Chinese character sets. To be able to search in these language natively the keywords need to be translated into the right characters. It is evident that especially for the methodology to search in online media, more unknown stimulants may be found when more languages are included. For example in China and Latin America where many new developments around stimulants have appeared in the last few years (INCB, 2020). In addition, other websites and databases that are more dedicated to publications on stimulants could be queried in addition to the MedISys search engine. A last limitation is that, currently, in both methodologies an expert must assess the results delivered by the systems. This is a time-consuming activity and preferably should be automated in the future.

4. Conclusions
We have shown that word embedding using scientific literature and text mining of the online media may both be used to detect new compounds that were unknown. Remarkably, both data sources and associated methodologies yielded different compounds, hence showing the complementary nature of the two sets of data and the necessity to analyse both the scientific literature and the online media. It is suggested that the developed approach can be used in other topics to find highly relevant but hitherto unknown for their (potential) use. This approach may in particularly be relevant for food safety authorities in their emerging risk identification activities to detect new compounds that may pose a health risk to consumers.

**Declarations**

**Acknowledgement**

The research leading to this result has received funding from the Ministry of Agriculture, Nature and Food Quality (LNV), the Netherlands via the Knowledge Base Programme (KB-23: Healthy and Safe Food for Healthy lives).

**References**

Altuncı, Y. A., Aydoğdu, M., Açikgöz, E., Güven, Ü., Düzağaç, F., Atasoy, A., Dağlıoğlu, N., & Annette Akgür, S. (2021). New Psychoactive Substance 5-MeO-MiPT In vivo Acute Toxicity and Hystotoxicological Study. *Balkan Med J, 38*(1), 34-42.

Bäckberg, M., Westerbergh, J., Beck, O., & Helander, A. (2016). Adverse events related to the new psychoactive substance 3-fluorophenmetrazine – results from the Swedish STRIDA project. *Clinical Toxicology, 54*(9), 819-825.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res., 3*(null), 1137–1155.

Benowitz, N. L., Jacob Iii, P., Mayan, H., & Denaro, C. (1995). Sympathomimetic effects of paraxanthine and caffeine in humans. *Clinical Pharmacology & Therapeutics, 58*(6), 684-691.

Bersani, F. S., Corazza, O., Albano, G., Valeriani, G., Santacroce, R., Bolzan Mariotti Posocco, F., Cinosi, E., Simonato, P., Martinotti, G., Bersani, G., & Schifano, F. (2014). 25C-NBOMe: preliminary data on pharmacology, psychoactive effects, and toxicity of a new potent and dangerous hallucinogenic drug. *BioMed research international, 2014, 734749.*

Biesterbos, J. W. H., Sijm, D. T. H. M., van Dam, R., & Mol, H. G. J. (2019). A health risk for consumers: the presence of adulterated food supplements in the Netherlands. *Food Additives & Contaminants: Part A, 36*(9), 1273-1288.
Bouzembrak, Y., Steen, B., Neslo, R., Linge, J., Mojtahehd, V., & Marvin, H. J. P. (2018). Development of food fraud media monitoring system based on text mining. *Food Control, 93*, 283-296.

Brandt, S. D., Walters, H. M., Partilla, J. S., Blough, B. E., Kavanagh, P. V., & Baumann, M. H. (2020). The psychoactive aminoalkylbenzofuran derivatives, 5-APB and 6-APB, mimic the effects of 3,4-methylenedioxymethamphetamine (MDA) on monoamine transmission in male rats. *Psychopharmacology (Berl), 237*(12), 3703-3714.

Carroll, B. C., McLaughlin, T. J., & Blake, D. R. (2006). Patterns and Knowledge of Nonmedical Use of Stimulants Among College Students. *Archives of Pediatrics & Adolescent Medicine, 160*(5), 481-485.

Chan, W. L., Wood, D. M., Hudson, S., & Dargan, P. I. (2013). Acute psychosis associated with recreational use of benzofuran 6-(2-aminopropyl)benzofuran (6-APB) and cannabis. *J Med Toxicol, 9*(3), 278-281.

Chowdhary, K. R. (2020). Natural Language Processing. In K. R. Chowdhary (Ed.), *Fundamentals of Artificial Intelligence* (pp. 603-649). New Delhi: Springer India.

Cinisi, E., Corazza, O., Santacroce, R., Lupi, M., Acciavatti, T., Martinotti, G., & di Giannantonio, M. (2014). New drugs on the Internet: the case of Camfetamine. *BioMed research international, 2014*, 419026-419026.

Coffin, V. L., & Spealman, R. D. (1989). Psychomotor-stimulant effects of 3-isobutyl-1-methylxanthine: comparison with caffeine and 7-(2-chloroethyl) theophylline. *Eur J Pharmacol, 170*(1-2), 35-40.

Cohen, P. A., Travis, J. C., Keizers, P. H. J., Deuster, P., & Venhuis, B. J. (2018). Four experimental stimulants found in sports and weight loss supplements: 2-amino-6-methylheptane (octodrine), 1,4-dimethylamylamine (1,4-DMAA), 1,3-dimethylamylamine (1,3-DMAA) and 1,3-dimethylbutylamine (1,3-DMBA). *Clin Toxicol (Phila), 56*(6), 421-426.

Corkery, J. M., Durkin, E., Elliott, S., Schifano, F., & Ghodse, A. H. (2012). The recreational tryptamine 5-MeO-DALT (N,N-diallyl-5-methoxytryptamine): A brief review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 39*(2), 259-262.

Corkery, J. M., Elliott, S., Schifano, F., Corazza, O., & Ghodse, A. H. (2012). 2-DPMP (desoxypipradrol, 2-benzhydrylpyrrolidine, 2-phenylmethylpiperidine) and D2PM (diphenyl-2-pyrrolidin-2-yl-methanol, diphenylprolinol): A preliminary review. *Prog Neuropsychopharmacol Biol Psychiatry, 39*(2), 253-258.

de Mello-Sampayo, C., Vaz, A. R., Henriques, S. C., Fernandes, A., Paradinha, F., Florindo, P., Faria, P., Moreira, R., Brites, D., & Lopes, A. (2021). Designer Cathinones N-Ethylhexedrone and Buphedrone Show Different In Vitro Neurotoxicity and Mice Behaviour Impairment. *Neurotox Res, 39*(2), 392-412.

Domagalska, E., Banaszkiewicz, L., Woźniak, M. K., Kata, M., Szpiech, B., & Kaliszan, M. (2020). Fatal N-Ethylhexedrone Intoxication. *Journal of Analytical Toxicology.*
ECD. (2020). WHO Expert Committee on Drug Dependence - TRS 1026 Forty-second report. In (pp. 44).

Fawzy, M., Wong-Morrow, W. S., Beaumont, A., & Farmer, C. K. T. (2017). Acute kidney injury and critical limb ischaemia associated with the use of the so called “legal high” 3-fluorophenmetrazine. CEN Case Reports, 6(2), 152-155.

Gatch, M. B., Dolan, S. B., & Forster, M. J. (2017). Locomotor and discriminative stimulus effects of four novel hallucinogens in rodents. Behav Pharmacol, 28(5), 375-385.

Helle, C., Sommer, A. K., Syversen, P. V., & Lauritzen, F. (2019). Doping substances in dietary supplements. Tidsskr Nor Laegeforen, 139(4).

Hill, S. L., Doris, T., Gurung, S., Katebe, S., Lomas, A., Dunn, M., Blain, P., & Thomas, S. H. (2013). Severe clinical toxicity associated with analytically confirmed recreational use of 25I-NBOMe: case series. Clin Toxicol (Phila), 51(6), 487-492.

INCB. (2020). Report of the International Narcotics Control Board for 2019. In. Vienne: UNITED NATIONS.

Joensuu, M., Tolmunen, T., Saarinen, P., Tiihonen, J., & Lehtonen, J. (2007). Reduced midbrain serotonin transporter availability in drug-naïve patients with depression measured by SERT-specific [123I] nor-²-CIT SPECT imaging. Psychiatry Research: Neuroimaging, 154, 125-131.

Katselou, M., Papoutsis, I., Nikolaou, P., Spiliopoulou, C., & Athanaselis, S. (2015). 5-(2-aminopropyl)indole: a new player in the drama of 'legal highs' alerts the community. Drug Alcohol Rev, 34(1), 51-57.

Kling-Petersen, T., Ljung, E., & Svensson, K. (1994). The preferential dopamine autoreceptor antagonist (+)-UH232 antagonizes the positive reinforcing effects of cocaine and d-amphetamine in the ICSS paradigm. Pharmacology Biochemistry and Behavior, 49(2), 345-351.

Končić, M. (2018). Getting More Than You Paid For: Unauthorized "Natural" Substances in Herbal Food Supplements on EU Market. Planta Med, 84(6-07), 394-406.

Lajtai, A., Mayer, M., Lakatos, Á., Kuzma, M., & Miseta, A. (2020). New psychoactive versus conventional stimulants - a ten-year review of casework in Hungary. Legal Medicine, 47, 101780.

Logan, B. K., Goldfogel, G., Hamilton, R., & Kuhlman, J. (2009). Five deaths resulting from abuse of dextromethorphan sold over the internet. J Anal Toxicol, 33(2), 99-103.

Majchrzak, M., Celiński, R., Kuś, P., Kowalska, T., & Sajewicz, M. (2018). The newest cathinone derivatives as designer drugs: an analytical and toxicological review. Forensic Toxicol, 36(1), 33-50.

Manier, S. K., Felske, C., Eckstein, N., & Meyer, M. R. (2020). The metabolic fate of two new psychoactive substances – 2-aminoindane and N-methyl-2-aminoindane – studied in vitro and in vivo to support drug testing. Drug Testing and Analysis, 12(1), 145-151.
Martin, S. J., Sherley, M., & McLeod, M. (2018). Adverse effects of sports supplements in men. *Australian prescriber, 41*(1), 10-13.

Marusich, J. A., Antonazzo, K. R., Blough, B. E., Brandt, S. D., Kavanagh, P. V., Partilla, J. S., & Baumann, M. H. (2016). The new psychoactive substances 5-(2-aminopropyl)indole (5-IT) and 6-(2-aminopropyl)indole (6-IT) interact with monoamine transporters in brain tissue. *Neuropharmacology, 101*, 68-75.

Marvin, H. J. P., Kleter, G. A., Frewer, L. J., Cope, S., Wentholt, M. T. A., & Rowe, G. (2009). A working procedure for identifying emerging food safety issues at an early stage: Implications for European and international risk management practices. *Food Control, 20*(4), 345-356.

Maskell, P. D., Smith, P. R., Cole, R., Hikin, L., & Morley, S. R. (2016). Seven fatalities associated with ethylphenidate. *Forensic Sci Int, 265*, 70-74.

Mayer, F. P., Burchardt, N. V., Decker, A. M., Partilla, J. S., Li, Y., McLaughlin, G., Kavanagh, P. V., Sandtner, W., Blough, B. E., Brandt, S. D., Baumann, M. H., & Sitte, H. H. (2018). Fluorinated phenmetrazine "legal highs" act as substrates for high-affinity monoamine transporters of the SLC6 family. *Neuropharmacology, 134*(Pt A), 149-157.

Meijer, N., Filter, M., Józwiak, Á., Willems, D., Frewer, L., Fischer, A., Liu, N., Bouzembrak, Y., Valentín, L., Fuhrmann, M., Mylord, T., Kerekes, K., Farkas, Z., Hadjidjeorgiou, E., Clark, B., Coles, D., Comber, R., Simpson, E., & Marvin, H. (2020). Determination and Metrics for Emerging Risks Identification DEMETER: Final Report. *EFSA Supporting Publications, 17*(7), 1889E.

Mestria, S., Odoardi, S., Federici, S., Bilel, S., Tirri, M., Marti, M., & Strano Rossi, S. (2020). Metabolism Study of N-Methyl 2-Aminoindane (NM2AI) and Determination of Metabolites in Biological Samples by LC–HRMS. *Journal of Analytical Toxicology*.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *https://arxiv.org/abs/1301.3781*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-751). Atlanta, Georgia: Association for Computational Linguistics.

Parks, C., McKeown, D., & Torrance, H. J. (2015). A review of ethylphenidate in deaths in east and west Scotland. *Forensic Sci Int, 257*, 203-208.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). Doha, Qatar: Association for Computational Linguistics.
Reissig, C. J., Carter, L. P., Johnson, M. W., Mintzer, M. Z., Klledinst, M. A., & Griffiths, R. R. (2012). High doses of dextromethorphan, an NMDA antagonist, produce effects similar to classic hallucinogens. *Psychopharmacology, 223*(1), 1-15.

Repke, D. B., Grotjahn, D. B., & Shulgin, A. T. (1985). Psychotomimetic N-methyl-N-isopropyltryptamines. Effects of variation of aromatic oxygen substituents. *Journal of Medicinal Chemistry, 28*(7), 892-896.

Richeval, C., Dumestre-Toulet, V., Wiart, J. F., Vanhoye, X., Humbert, L., Nachon-Phanithavong, M., Allorge, D., & Gaulier, J. M. (2019). New psychoactive substances in oral fluid of drivers around a music festival in south-west France in 2017. *Forensic Sci Int, 297*, 265-269.

Roque Bravo, R., Carmo, H., Silva, J. P., Valente, M. J., Carvalho, F., Bastos, M. d. L., & Dias da Silva, D. (2020). Emerging club drugs: 5-(2-aminopropyl)benzofuran (5-APB) is more toxic than its isomer 6-(2-aminopropyl)benzofuran (6-APB) in hepatocyte cellular models. *Archives of Toxicology, 94*(2), 609-629.

Rothman, R. B., Partilla, J. S., Baumann, M. H., Lightfoot-Siordia, C., & Blough, B. E. (2012). Studies of the biogenic amine transporters. 14. Identification of low-efficacy "partial" substrates for the biogenic amine transporters. *The Journal of pharmacology and experimental therapeutics, 341*(1), 251-262.

Schwartz, A. R., Pizon, A. F., & Brooks, D. E. (2008). Dextromethorphan-induced serotonin syndrome. *Clin Toxicol (Phila), 46*(8), 771-773.

Shafi, A., Berry, A. J., Sumnall, H., Wood, D. M., & Tracy, D. K. (2020). New psychoactive substances: a review and updates. *Therapeutic advances in psychopharmacology, 10*, 2045125320967197-2045125320967197.

Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F. B., & Wattenberg, M. (2016). Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*.

Tolliver, B. K., & Carney, J. M. (1995). Locomotor stimulant effects of cocaine and novel cocaine analogs in DBA/2J and C57BL/6J inbred mice. *Pharmacol Biochem Behav, 50*(2), 163-169.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95-98.

Welter, J., Kavanagh, P., Meyer, M. R., & Maurer, H. H. (2015). Benzofuran analogues of amphetamine and methamphetamine: studies on the metabolism and toxicological analysis of 5-APB and 5-MAPB in urine and plasma using GC-MS and LC-(HR)-MS(n) techniques. *Anal Bioanal Chem, 407*(5), 1371-1388.

Wohlfarth, A., Roman, M., Andersson, M., Kugelberg, F. C., Diao, X., Carlier, J., Eriksson, C., Wu, X., Konradsson, P., Josefsson, M., Huestis, M. A., & Kronstrand, R. (2017). 25C-NBOMe and 25I-NBOMe metabolite studies in human hepatocytes, in vivo mouse and human urine with high-resolution mass spectrometry. *Drug Test Anal, 9*(5), 680-698.
Figures

Figure 1

The word embeddings from the trained model projected in three-dimensional space centralizing the word embedding for “stimulant”. A darker color represents a denser cluster of neighbors. Examples of the neighboring words are plotted next to their corresponding points in space. The projection was created with t-distributed stochastic neighbor embedding (t-SNE) using cosine distance, a perplexity of 30, a learning rate of 10 and 1000 iterations with the Tensorflow embedding projector (Smilkov, et al., 2016).
Figure 2

The word embeddings of the word “stimulant” and its closest 50 neighbors taken from the trained word embedding model projected in two-dimensional space. The projection was created with t-distributed stochastic neighbor embedding (t-SNE) using cosine distance, a perplexity of 5, a learning rate of 10 and 5000 iterations with the Tensorflow embedding projector (Smilkov, et al., 2016).
Figure 3

Dashboard showing, in various panels, information of the collected articles. Panel are numbered from left to right, top to bottom. Panel 1: Total number of articles collected. Panel 2: Number of articles per week. Panel 3: Number of articles collected with a defined keyword. Panel 4: World map showing the number of articles per country. Panel 5: Table with the hyperlink to the original location of the articles. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 4

Network visualisation of the titles and abstracts. The network was created with VOSviewer, only the top 50 terms that were mentioned at least 13 times are shown.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf