Use of literature mining for early identification of emerging contaminants in freshwater resources

Hartmann, Julia; Wuijts, Susanne; van der Hoek, Jan Peter; de Roda Husman, Ana Maria

DOI
10.1186/s13750-019-0177-z

Publication date
2019

Document Version
Final published version

Published in
Environmental Evidence

Environmental Evidence

METHODOLOGY

Open Access

Use of literature mining for early identification of emerging contaminants in freshwater resources

Julia Hartmann[1,2*], Susanne Wuijts[1,3], Jan Peter van der Hoek[2,4] and Ana Maria de Roda Husman[1,5]

Abstract

Chemical and microbial contaminants in the aquatic environment pose a potential threat to humans and to ecosystems. Humans may be exposed to contaminants in water resources when used for drinking water production, agriculture, aquaculture or recreation. Climatological, social and demographic changes, as well as the increasing sensitivity of analytical techniques, may result in the augmented detection of contaminants. Recent research has shown that it takes about 15 years from the time of the first scientific study mentioning the presence of a contaminant in the environment for the issue to peak in scientific attention and regulatory action. One possible factor influencing this lengthy period is that the first article becomes lost in the vast number of publications. In this study, we therefore developed a methodology using literature mining to identify the first scientific study which reports the presence of a contaminant in the aquatic environment. The developed semi-automated methodology enables health and environment agencies to inform policy makers about contaminants in the aquatic environment that could be significant for public and environmental health in national, international and river basin settings. The methodology thereby assists the proactive governance of emerging contaminants in the aquatic environment. This was illustrated by a retrospective analysis of the period of emergence in the Netherlands of: (1) perfluorooctanoic acid in surface water, and (2) biological industrial wastewater treatment systems as potential infection sources for Legionnaires´ disease.

Keywords: Chemical, Microbial, Emerging pathogen, Text mining, R-based, Early warning, Water, Health, Emergence

Background

Human activities result in the release of contaminants into the aquatic environment. Anthropogenic sources contaminating the aquatic environment include the effluents of municipal wastewater treatment plants (WWTPs), industrial wastewater discharges, as well as runoff from agricultural land and urban areas [1]. Moreover, demographic, social and climatological changes aggravate the impact of human activities on the aquatic environment. Examples of these changes are the increased volumes and changed composition of wastewater caused by urbanisation and the decreasing dilution capacities of receiving water bodies due to droughts which results in higher concentrations of contaminants in water bodies [2, 3]. The increasing sensitivity of analytical techniques also enables the augmented detection of contaminants in the aquatic environment [3, 4].

Anthropogenic contamination may contain both chemical and microbial contaminants. For instance, the effluent of municipal WWTPs, despite advanced treatment steps, may contain pharmaceutical and personal care products [5], antibiotic resistant bacteria [6] and antibiotic resistance genes [7]. Also, industrial wastewaters, dependent on the type of industry, have been found to contain several chemical contaminants, such as dyes, solvents and catalysts [8]. Microbial contaminants have also been detected in industrial wastewater, for instance viruses that have been accidently released during vaccine production [9]. Chemical and microbial contaminants released into the aquatic environment can not only pose a threat to human health when water resources are

*Correspondence: Julia.hartmann@rivm.nl
[1] National Institute for Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, The Netherlands
Full list of author information is available at the end of the article

used for drinking water production or recreation, but can also impact aquatic organisms. In this study, we refer to emerging contaminants for which the threat posed to human health or the aquatic environment is still unclear.

In a recent study, we showed that the current risk governance of contaminants in the aquatic environment can be improved by the more timely identification of contaminants which are of potential concern [10]. In that study, we analysed the current policy on the risk governance of emerging contaminants in the aquatic environment in the Netherlands, Germany, Switzerland and the state of Minnesota and found that timely identification enabled, among other things, appropriate risk management strategies. Furthermore, Halden [11] investigated, in retrospect, the association between the number of scientific publications about certain chemical environmental contaminants, such as dichlorodiphenyltrichloroethane (DDT) and 1,4-dioxane, and the regulatory actions subsequently taken. He found that it generally took about 15 years from the first scientific publication about a contaminant to a peak in number of scientific publications. The peak in scientific attention was found, in many cases, to be associated with regulatory or mitigation actions. The period from the first scientific publication being released to the time at which it reaches the peak of scientific attention is referred to as the 'period of emergence of concern' by Halden [11]. Shortening the period of emergence of concern may accelerate the introduction of regulatory actions to control chemical contaminants in the environment and thus limit environmental effects.

Although Halden [11] looked specifically at the emergence of concern about chemical contaminants, similar trends can be found for emerging microbial contaminants. Specific pathogens have (in retrospect) been shown to be present in the environment and linked to human sources long before the disease that they cause had gained attention [12]. For the Aichi Virus this has been illustrated by Lodder et al. [13]. The Aicihi virus was reported in humans for the first time in 1989. However, Lodder et al. [13] analysed environmental water samples from the Netherlands from 1987 and found that the Aichi virus had been circulating in the Dutch population well before its initial detection in humans. The fact that the Aichi virus was identified in water samples showed that the virus was already present in humans in 1987; otherwise it could not have been detected in the aquatic environment. Furthermore, the properties that cause concern among scientists and regulators about contaminants in the aquatic environment, especially when used for the production of drinking water, are similar for chemical and microbial contaminants. These properties include pathogenicity or toxicity, persistence and mobility [14, 15]. Therefore, decreasing the period of the emergence of concern about microbial contaminants is also important if timely mitigation actions are to be ensured.

Currently, we believe that the first scientific article about the presence of a contaminant in the aquatic environment is not picked up by regulators due to the large number of publications. It is not until more articles are published about the specific contaminant that the signal about the presence of the contaminant in the environment is picked up by regulators, as is shown by Halden [11]. We hypothesise that the period of emergence of concern about contaminants can be reduced by the systematic search of the universal scientific literature for articles reporting the first detection of a contaminant in the aquatic environment. As many articles about contaminants in the aquatic environment are published every day, the manual analysis of the scientific literature would be too complex, subjective and time consuming.

Text mining can be used to automate some parts of systematic literature reviews. The term refers to the automated extraction of (parts of) articles that are relevant to the researcher, or to the data mining of articles, which enables associations to be found between parts of texts [16, 17]. Text mining has been shown useful in biomedical research for several applications, such as in the identification of eligible studies and the allocation of a list of genes to inform on their role in diseases [18]. Here, eligible studies refer to articles reporting on original research that is considered relevant to the scope of the systematic literature review. Others in the field of evidence-based software engineering for systematic literature reviews have used the term "primary studies" for this purpose [19]. Furthermore, Van de Brug et al. [20] have used text mining to devise an early warning mechanism to detect potential food related risks. Sjerps et al. [21] have also used text mining to identify signals of potential emerging chemical risks to drinking water quality by combining search terms connected to chemical contaminants and the aquatic environment. However, this approach did not include microbial contaminants and was not specifically aimed at generating first reports on the presence of contaminants in the aquatic environment.

Over the past years, several software tools have been developed which integrate text mining in the systematic literature review process [22]. In this study, we assessed the applicability of two such tools, namely the StArt Tool and Adjutant. The StArt Tool automates the eligible study selection process by scoring articles based on the number of occurrences of the search terms in the title, abstract and keywords (open source and available at http://lapes.dc.ufscar.br/tools/start_tool, automates) [22]. The rationale of the StArt tool is that the highest scoring articles are most relevant to the performed search and should thus be selected as eligible studies.

Adjutant, another software tool, can be used to query the PubMed® database and perform unsupervised clustering on the retrieved collection of articles [23]. Adjutant is available from https://github.com/amcrisan/Adjutant. In this study, we assessed the applicability of two software tools, namely the StArt Tool and Adjutant, to identify articles that report on the detection of a contaminant in the aquatic environment for the first time.

The objective of this study is to introduce a methodology using literature mining to identify the first signal of the detection of a chemical or microbial contaminant in the aquatic environment. To keep the search as concise as possible, we focus in this study on freshwater resources. First, the development of the methodology is explained making use of the selected software tools ("Methodology development" section). Then, the application of the developed methodology to recent scientific literature is shown ("Results of applying methodology to recent literature" section). Finally, a retrospective validation of the proposed methodology is discussed using the period of emergence of concern in the Netherlands of (1) perfluorooctanoic acid (PFOA) in surface water and (2) biological industrial wastewater treatment systems as potential infection sources of Legionnaires´ disease ("Retrospective validation of the developed methodology" section).

The developed methodology adds to evidence synthesis by combining signals of first detections of contaminants in the aquatic environment into manageable information. Health or environment agencies can use the methodology to inform policy makers about signals of emerging contaminants in the aquatic environment that could be relevant for public or environmental health in a national, international or river basin setting. The methodology thereby assists the proactive governance of emerging contaminants in the aquatic environment and contributes to the objective and proactive use of scientific evidence to inform policy makers.

Methodology development

A systematic literature review has three phases: planning, conducting and reporting. The planning phase includes identifying the need for a review and creating a review protocol. In the conducting phase, authors search for literature, identify and appraise eligible studies, and extract and synthesise data. In the final phase the results of the review are reported to relevant communities [19]. In this study, we have used R-based coding in the conducting phase to make the review process more efficient. A graphical representation of the development of the methodology is shown in Fig. 1 and is described in this section. The reporting phase is not automated by the developed methodology because, in this study, the reporting phase includes the elucidation of the relevance of the identified contaminants in a national, international or river basin setting.

In this study, the first signal of the detection of a chemical or microbial contaminant in the aquatic environment refers to a scientific article. In order to find this article, we use text mining of scientific articles, from now on referred to as literature mining. Here, literature mining is the automated textual analysis of the combination of 'title' and 'abstract'. This does not include the analysis of the data sets produced by the different articles [24]. The developed methodology is therefore applicable to all scientific literature, also when the full-text of the article cannot be accessed. The methodology is written in R-studio, available at https://www.r-project.org/ to make it freely accessible. All codes written in R referred to in the following methodology are added as supplementary material in Additional file 1.

The planning phase

The review protocol was designed so that scientific articles that report on the first identification of chemical or microbial contaminants in the aquatic environment could be found. The search was conducted in Elsevier's Scopus®, the largest abstract and citation database of peer-reviewed literature worldwide [25]. In order to find articles reporting on the first identification of contaminants in the aquatic environment, relevant search terms and inclusion and exclusion criteria were defined.

Search query

The search terms used in the review are shown in Fig. 2. The search query itself was a combination of four concepts, namely *contaminant*, *detection*, *new*, and *aquatic environment*. In order to keep the search query as specific as possible, it was decided to focus on freshwater resources. Each concept included several synonyms and was searched for in the title, abstract and keywords. The search query was set up using expert opinion and a list of fourteen a priori selected articles (see Table 1). The fourteen articles report the identification of chemical or microbial contaminants in the aquatic environment for the first time and could thus be used to test the effectiveness of the proposed methodology. The articles were found using a simple search in Google Scholar® using the search terms "first" and "detect* OR identif*". Furthermore, articles which the authors came across in previous research and that reported on the first identification of chemical or microbial contaminants in the aquatic environment were also included in Table 1.

Experts from different backgrounds, such as chemistry, microbiology, and hydrology, also provided input and feedback on a list of search terms using an iterative approach, thus ensuring that a comprehensive list of

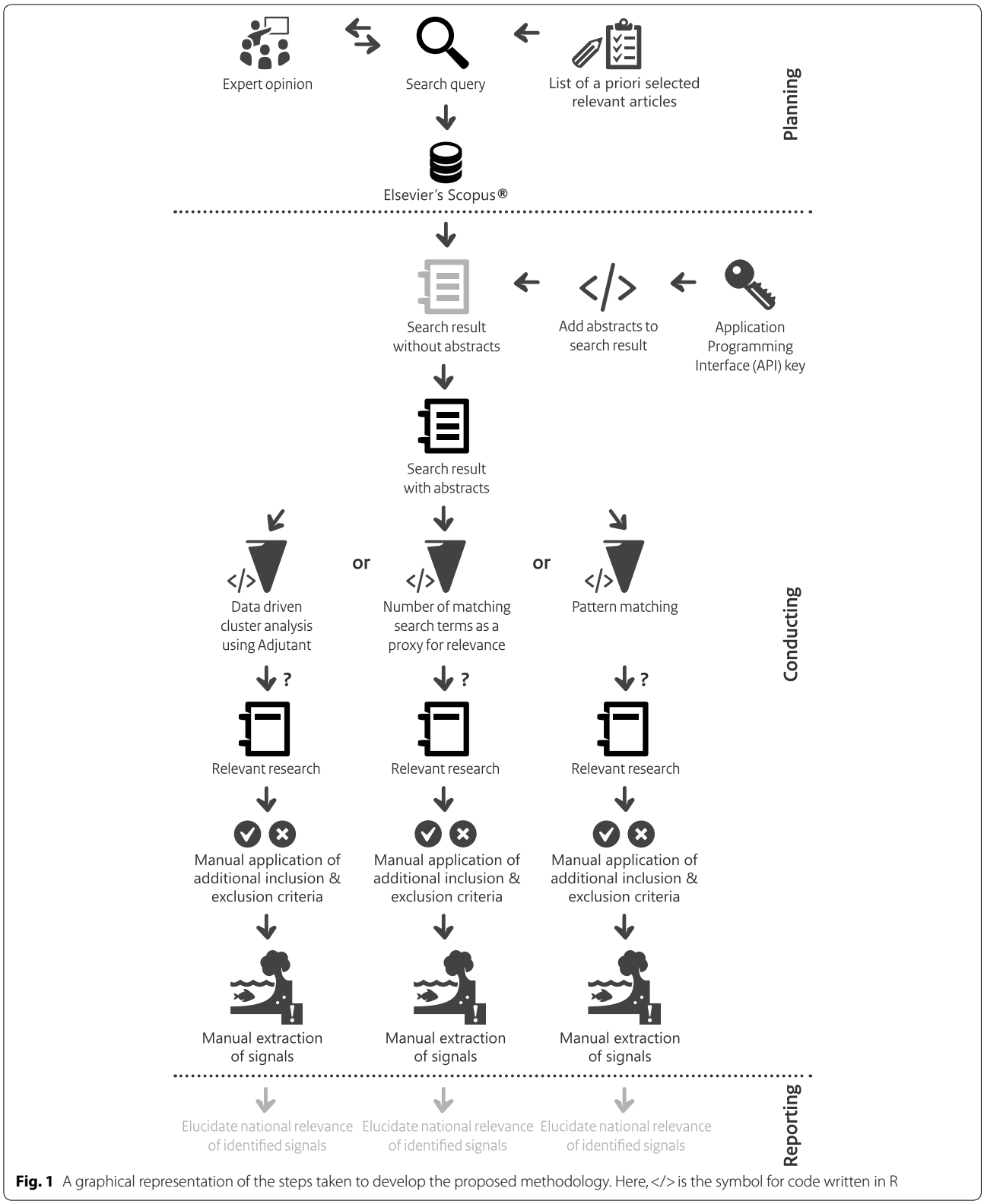

Fig. 1 A graphical representation of the steps taken to develop the proposed methodology. Here, </> is the symbol for code written in R

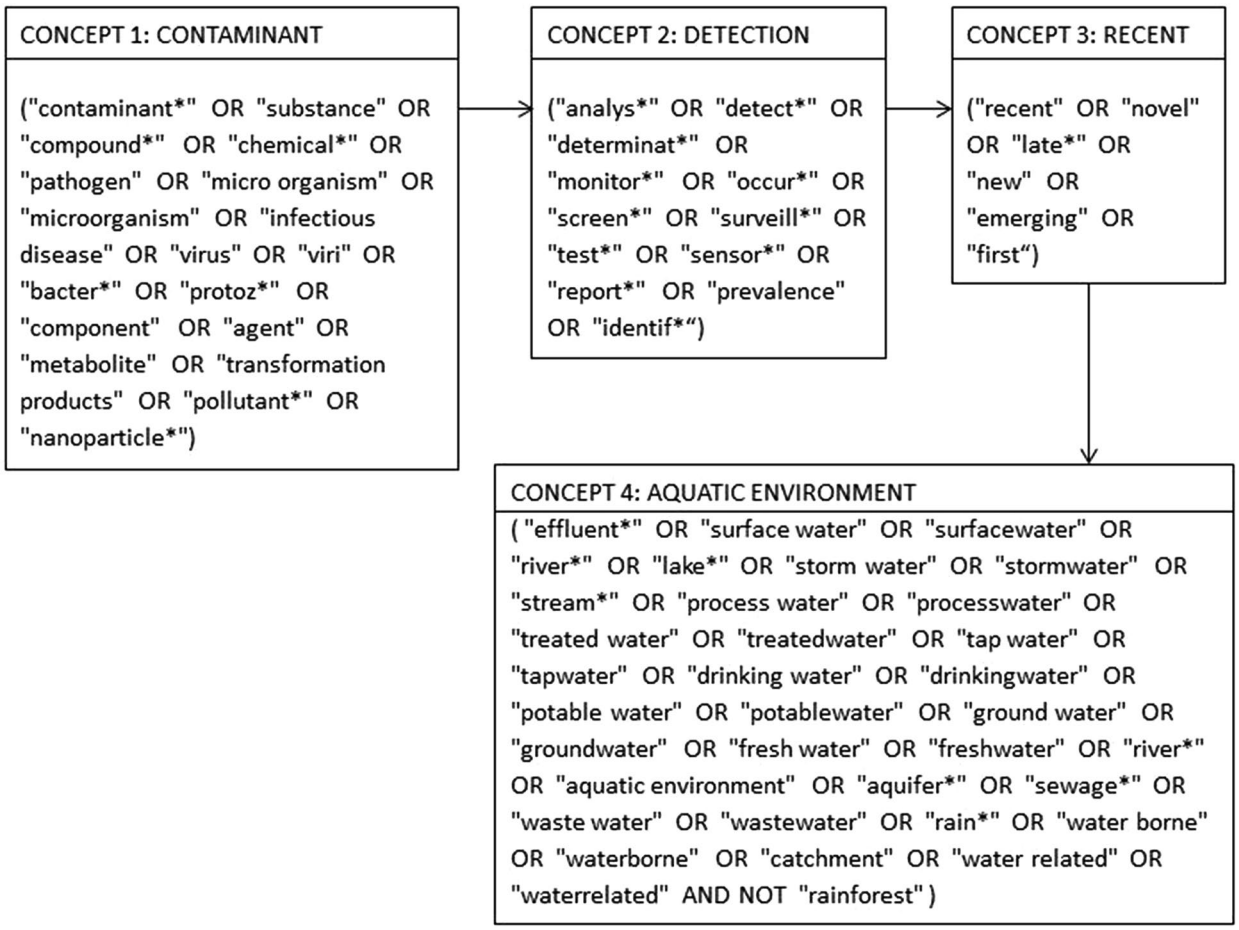

Fig. 2 Search terms used to search Scopus® for articles reporting on the first identification of chemical or microbial contaminants in the aquatic environment. Search terms were searched for in title, keywords and abstracts. Additional information: _ = search term was used with, and without, the use of a space, * = any combination of characters, → = AND

search terms was obtained. In order to keep the search query as concise as possible, it was decided that a number of specific kinds of contaminants would not be included in concept 1 (e.g. pharmaceuticals, pesticides or *E. coli*). However, we did add the term 'nanoparticle' as nanoparticles are not always referred to as compounds or contaminants and records referring to nanoparticles would otherwise be missed by the presented methodology.

Inclusion and exclusion criteria

In the query in Scopus we limited the search to scientific articles, reviews and articles in press written in English. Although we were looking for original research, reviews were also included as authors of original research might not have been aware that they had identified a contaminant for the first time, but a reviewer might have picked up on it. Furthermore, the search query excluded records from the following subject areas: economics, econometrics and finance, business, management and accounting, dentistry, and psychology. Finally, to develop the methodology, only articles published between 2006 and 2012 were included, as the set of articles retrieved with the search query had to contain the a priori selected articles (see Table 1, publication year 2006–2012).

Some inclusion and exclusion criteria could not be included in the search query, but were used to manually select eligible studies in the conducting phase. Although interesting, studies about new analytical techniques, new bio indicators, new toxicity results for known contaminants, new detections in the marine environment and in soil, and new removal techniques for known contaminants, were outside the scope of this study and not considered eligible studies. Studies about new detections in aquatic biota and aquatic plants were included as these are direct signals of aquatic contamination. However, first detections in terrestrial plants

Table 1 List of 14 a priori selected articles that report on the identification of specific contaminants in the aquatic environment for the first time

References	Publication year	First detection of	Detection in	Cluster	Sorted on position	Total search terms
Sultan and Gabryelski [28]	2006	Several contaminants, most intriguing Glycolic acid	Drinking water	Not-clustered	11	17
Terasaki et al. [29]	2008	Five aryl hydrocarbons, including a novel chlorinated aryl ether	Surface water (port where industrial effluent from paper mill is discharged)	Not-clustered	241	12
Conley et al. [30]	2008	Lovastatin	Surface water	Not-clustered	1491	9
Radjenović et al. [31]	2008	Biodegradation product of β-blocker atenolol	Wastewater samples	Not-clustered	41	15
Conley et al. [27]	2008	Norfluoxetine	Surface water	N/A	N/A	N/A
Xiao et al. [32]	2008	Gatifloxacin	River, influent and effluent of sewage treatment plant	Not-clustered	2799	8
Söderström et al. [33]	2009	Oseltamivir	Surface water	Pharmaceut-concentr	1672	9
Hamza et al. [34]	2009	Human bocavirus	Surface water (river)	Infect-diseas	3047	8
Ferrer and Thurman [35]	2010	Lamotrigine and glucuronide	Wastewater and drinking water samples	Sampl-detect	3	19
Farré et al. [36]	2010	C60 and C70 fullerenes and N-Methylfulleropyrrolidine C60	Effluent wastewater treatment plant	Inf-chemic-concentr	162	13
Zhao et al. [37]	2010	Three new disinfection by-products (DBPs): 2,6-dichloro-3-methyl-1,4-benzoquinone, 2,3,6-trichloro-1,4-benzoquinone, and 2,6-dibromo-1,4-benzoquinone	Drinking water	Chlorin-disinfect	997	10
Kleywegt et al. [38]	2011	Roxithromycin and enrofloxacin	Drinking water resource	Pharmaceut-concentr	327	12
Pereira Rde et al. [39]	2011	Identification of new ozonation disinfection by-products of 17β-estradiol	Groundwater	Ozon-effect	14,570	5
Su et al. [40]	2012	Three gene cassette arrays, *aac(6′)-Ib-cr-aar-3-dfrA27-aadA16*, *aacA4-catB3-dfrA1* and *aadA2-lnuF*	Surface water	Resist-antibiot	3904	8

The 5th column, named 'Cluster', is the result of the data driven cluster analysis using Adjutant, which is explained in "Eligible study selection approaches" section. The 6th and 7th column, named 'Sorted on position' and 'Total search terms' show the ranking of the a priori selected articles based on presence of search terms, this approach is also explained in "The conducting phase" section. *N/A* not applicable (Conley et al. [27] was not found with the search query used)

were not included as eligible studies. Articles about drinking water or wastewater treatment techniques were excluded as the aim of the developed methodology was to identify first detections of contaminants in the aquatic environment and not to identify new treatment techniques used to treat contaminated water. However, articles reporting the first identification of contaminants created during treatment, e.g. newly identified disinfection by-products, were included.

An overview of the search query and the inclusion and exclusion criteria used is shown in Additional file 2.

The conducting phase

The search query (shown in Additional file 2) was used to search Scopus®; this generated 27,516 articles. As Scopus® does not have the functionality to export more than 2000 records, including all bibliographic information, R-based coding was used to add abstract

information to each record using the *Rscopus* package (see Fig. 1) [26]. In order to retrieve abstract information from Scopus® by using R, an Application Programming Interface (API) key is needed which can be requested from Elsevier, using this link https://dev.elsevier.com/. The full script for this step of the methodology can be found in Additional file 1.

After the code was run, the list of 27,516 articles contained abstract information. It was found that only 13 of the 14 a priori selected articles were included in this dataset. Conley et al. [27] was not found by the search query shown in Additional file 2. This is due to the fact that the first detection of the contaminant was not mentioned in the title or abstract. We continued developing the methodology with the other thirteen articles shown in Table 1.

The following step in a review process would be to manually select eligible studies based on title and abstracts. However, the high number of records makes the manual selection of eligible studies unrealistic, so R was used to automate the eligible study selection process.

Eligible study selection approaches

Available software tools were used to automate the eligible study selection process in this research, namely the StArt tool [22] and Adjutant [23] (see also Fig. 1). As the StArt tool was not R-based we implemented the rationale used in the StArt tool in R. Adjutant could be directly used in R. We also assessed whether available text mining functionalities within R could be used. An explanation of the three approaches follows below (see also Fig. 1). Each approach has been computed into a separate R-based code which can be found in Additional file 1.

1. Data driven cluster analysis using Adjutant: Adjutant was originally developed to cluster articles retrieved from the Pubmed database [23]. With minor adjustments to the package, Adjutant turned out to be useful for Scopus® data as well. Furthermore, the package uses 'stopwords', which are words that are considered to be so widely used in the collection of articles that they are irrelevant to the content clustering analysis. We added additional stopwords to the package based on our search query, namely: water, study, studies, studied, species, region, and stable. These words were chosen because they are widely present in the set of articles exported from Scopus.
2. Number of search terms as a proxy for relevance: the rationale of the StArt tool (as discussed in "Background" section) was used as a guide for working out how to automatically identify eligible studies using R [19, 28]. The developers of the StArt tool advise using different values for occurrences in different parts of the text, especially lower values for occurrences in keywords. Occurrences of search terms in keywords should be rated lower because keywords are often not exported from search databases into the StArt tool. Also, as authors are obliged to choose a limited number of keywords, they might not be able to catch the research subject in this limited number [19]. We did not have any information on the keywords, as these were not in the dataset we exported from Scopus®. Therefore, we examined whether specific terms from the search query were more frequent in the a priori selected articles than others. In that way, we were able to add more weight to those relevant terms when scoring articles. This was done using the *tm* and *quanteda* packages in R [29, 30].
3. Pattern matching: the abstracts of the fourteen a priori selected articles (see Table 1) were assessed so that we could find a common pattern which would indicate the relevance of these articles to the present study. First, the abstract and titles were split into sentences and then the pattern, shown in Additional file 1, was used to select relevant articles using string pattern matching. In Additional file 1, it is shown that the pattern checks out for a combination of different word stems (e.g. 'new' and 'detect') in one sentence. However, these do not need to occur next to each other, hence the addition of 0–70 characters between the word stems. This is different from the search query used in Scopus®, as Scopus® is unable to search for specific combinations of words or word stems in one sentence. Also, by using the pattern matching in R, the matching sentence can be retrieved from the specific abstract which makes analysis less time consuming.

The applicability of the three approaches to automate the eligible study selection process was analysed using the fourteen a priori selected articles. However, one of these fourteen articles was not found in any of the approaches [27]. The first approach, namely data driven cluster analysis using Adjutant (Script 2), resulted in 48 clusters. However, 12,959 records (53%) were not clustered. Figure 3 shows the clusters that have been constructed and Table 1 shows the clusters in which the a priori selected records were sorted by Adjutant. Five of the a priori selected records were not clustered. Also, the eight records that were clustered, were divided over six different clusters. Therefore, there was no clear indication as to which of the clusters contained relevant information on the first detection of contaminants in the aquatic environment. Thus, data driven cluster analysis using Adjutant was not considered a feasible

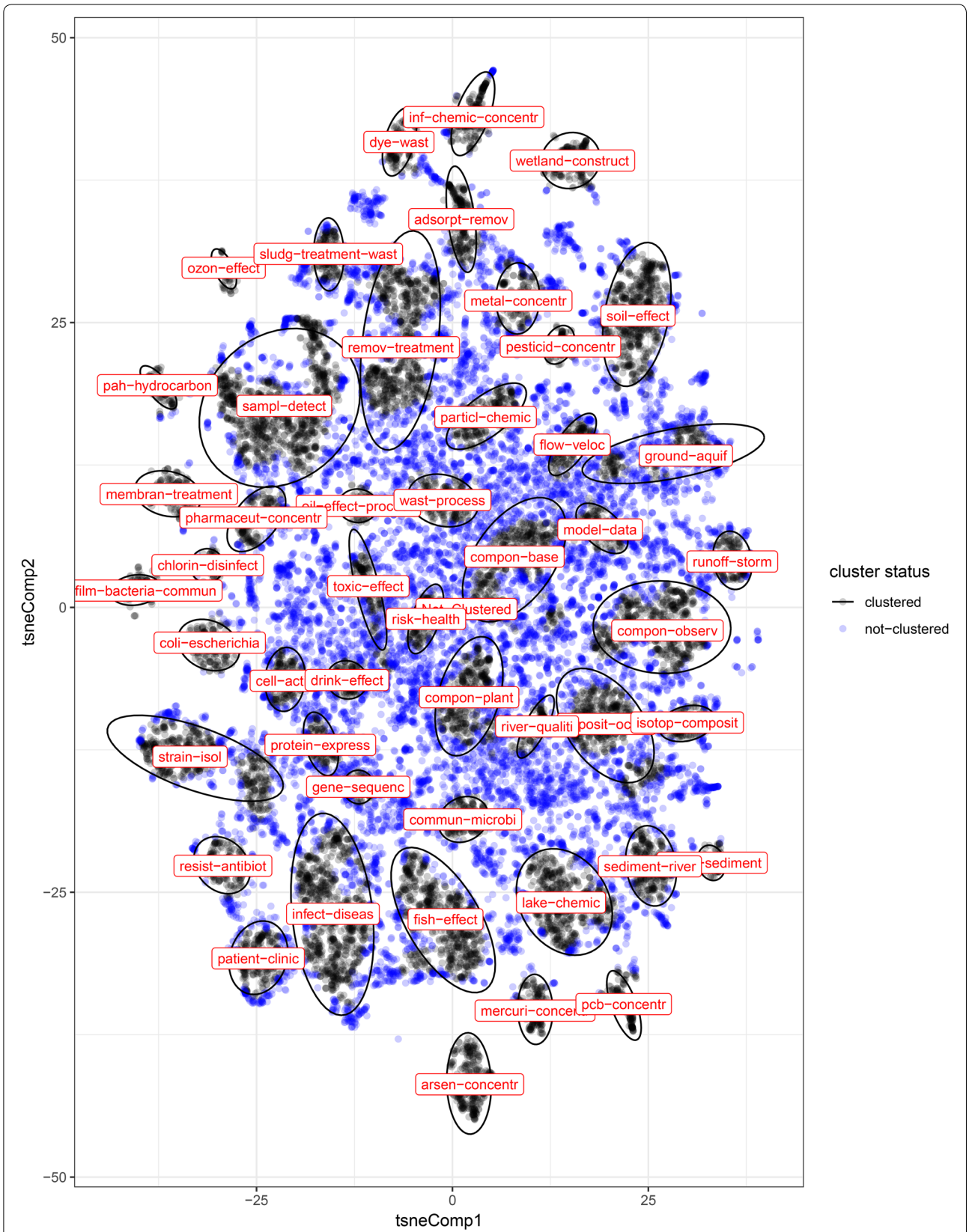

Fig. 3 Result of the data driven cluster analysis using the Adjutant package (Script 2). The names of the clusters are the two most commonly used word stems in the specific cluster

approach for the automation of the eligible study selection process in this research.

The second approach to automate the eligible study selection process that was assessed was based on the classification approach used in the StArt tool [19, 28]. Figure 4 shows the most used search terms in 13 of the a priori selected articles (Conley et al. [27] was not found by the search query used). There is no clear indication which of the concepts (see "Search query") is most distinguishably present in these relevant articles. Therefore, the records were sorted based on the presence of all the search terms using the *quanteda* package, with no additional weights added to any concepts or search terms. Table 1 shows that not all a priori selected articles are ranked high. Therefore, the ranking of articles that was based on the frequency of search terms was found not to be applicable to automate the eligible selection process in this study.

The third approach assessed to automate the eligible selection process was pattern matching. The dataset contained 4299 records that matched the pattern based on the a priori selected articles. This is 15.6 percent of the original number of records exported from Scopus®. All but one, namely Conley et al. [27], of the a priori selected articles were included in the 4299 records.

Because the pattern matching approach was the only approach that clustered the a priori selected articles together, we found pattern matching to be the best approach to automate the eligible study selection in this research. Using this approach the eligible study selection process is not yet fully automated as the list of matched records still needs to be manually checked. However, the number of records that is likely to include most eligible articles and thus should be prioritised for manual checking was decreased by almost 85 percent. Therefore, pattern matching was chosen as the approach to automate (part) of the screening process.

Sensitivity and specificity analysis

A sensitivity and specificity analysis of the developed pattern was performed using the fraction true or false negatives and true or false positives. Here, false positives are articles that did not report the first detection of a contaminant in the aquatic environment but were extracted as eligible studies using the pattern defined in Additional file 1. False negatives are articles which did not match the pattern although these articles reported on the first detection of a contaminant in the aquatic environment. Often in computational linguistics, the focus is on the proportion of true and false positives recalled by the methodology, since no information is available on the documents that were not retrieved by the methodology [31]. However, here we have information on the articles that were eliminated using the pattern defined in Additional file 1. Therefore, we used the definitions of sensitivity and specificity as shown in Eqs. 1 and 2 following the Receiver Operating Characteristics (ROC) analysis [32].

$$sensitivity = \frac{fraction\ of\ true\ positives}{fraction\ of\ true\ positives + fraction\ of\ false\ negatives} \quad (1)$$

$$specificity = \frac{fraction\ of\ true\ negatives}{fraction\ of\ true\ negatives + fraction\ of\ false\ positives} \quad (2)$$

Results of applying methodology to recent literature

In this section, the results of applying the developed methodology, as explained in ("Methodology development" section, to recent literature, namely articles published between 2016 and 27th of August 2018, are presented. Running the search query shown in Additional file 2, adjusted to the new time period, resulted in 22,570 articles being found in Scopus®. A list containing these records was exported from Scopus® and the code to add abstract information (see "The conducting phase" section) was used. Pattern matching was run to identify eligible studies, which resulted in 3650 records (16.0 percent of the original dataset) containing 3983 sentences that matched the pattern. These records were exported to an excel file that contained the articles' Electronic Identifier (EID), authors, title, publication year, journal, volume, page information, citations, Digital Object Identifier (DOI), link to the article in Scopus®, abstract and the sentence that matched the pattern.

Then, eligible studies were again selected by applying additional criteria to the remaining dataset of 3650 articles. The inclusion and exclusion criteria defined in "Inclusion and exclusion criteria" section were used. After manual analysis, 359 articles were selected as eligible studies, as shown in Additional file 3. The contaminants detected for the first time in these studies were categorised manually as chemical or microbial.

Of the 359 articles, 173 were on chemical contaminants and 186 on microbial contaminants. The next step would

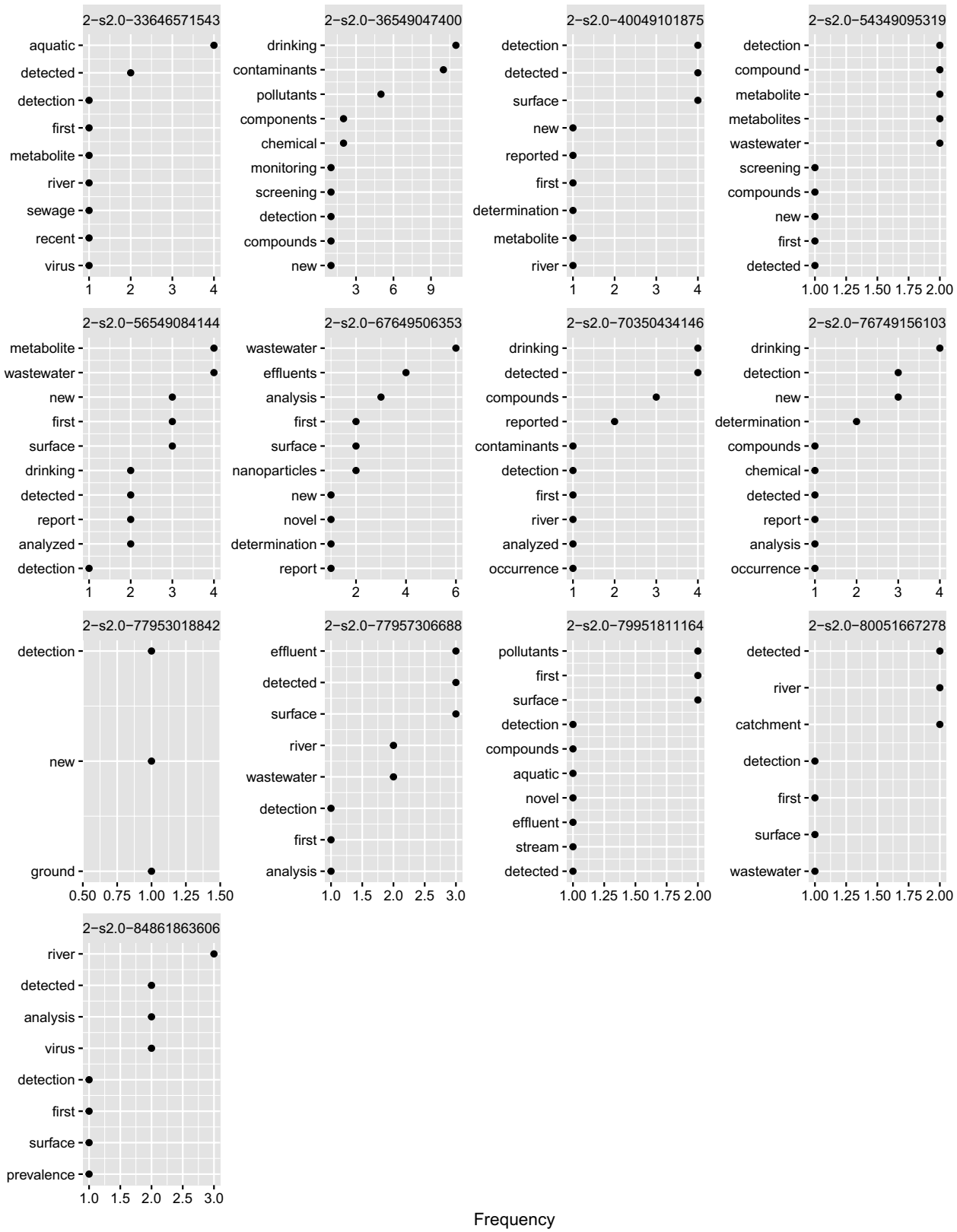

Fig. 4 Overview of the search terms that were used most often in thirteen of the a priori selected relevant articles

be to identify the relevance of the contaminants identified for the first time as potential threats to public and environmental health in national, international or river basin settings. The elucidation process is not automated by the developed methodology and therefore not within the scope of this study. However, we are planning to further develop the elucidation process in detail in future research.

Results of the sensitivity and specificity analysis

In order to find the fraction of false and true negatives, we analysed a random selection of 1750 articles from the 23,217 articles (published between 2006 and 2012) that did not match the pattern. We found that 32 of the 1750 articles did report on the first detection of a contaminant in the aquatic environment, resulting in a fraction of true and false negatives of 0.982 and 0.018, respectively. The results of the analysis are shown in Additional file 4. Out of the 3650 articles extracted as eligible studies, 359 articles were true positives resulting in a fraction of true and false positives of 0.098 and 0.902, respectively. Therefore, using Eqs. 1 and 2, a sensitivity of 84.5% and a specificity of 52.1% were found.

Retrospective validation of the developed methodology

Could the developed methodology have contributed to the earlier identification of any of today´s emerging contaminants in the aquatic environment? To answer this question, we further analysed two examples of contaminants, one chemical and one microbial, which have caused great concern over the past years. We ran the methodology as defined above and assessed whether the use of the proposed text mining methodology would have decreased the period of emergence of concern in the Netherlands. The chemical contaminant used as an example was perfluorooctanoic acid (PFOA), which is an anthropogenic chemical belonging to the group of per- and polyfluoroalkyl substances (PFASs) [33]. The microbial contaminant example was the family of the *Legionella* bacteria.

Perfluorooctanoic acid (PFOA)

Since the 1940s, PFOA has been used in many industrial applications, for instance in the production of Teflon®. In 1978, it was first established that PFOA induces immunotoxicity and other adverse effects in monkeys. However, Grandjean and Clapp [34] showed that this, and other early toxicity information, was not published or was overlooked. Regulatory actions were, therefore, only initiated after the analysis of blood serum samples taken in 2000 revealed that PFOS and PFOA were detectable in all Americans [35]. In 2010, the major PFOA producing company in the United States of America stated that it had decreased its PFOA emissions by 95 percent [34].

In the Netherlands, Dupont had been using PFOA since 1970 to produce Teflon and had replaced it voluntarily in 2012 by a different perfluorinated compound. In 2015, groundwater which had been used for the production of drinking water was investigated for possible contaminants and found to be polluted by PFOA as the result of industrial wastewater discharges and subsequent infiltration into groundwater in the period of 1970–2012 [36, 37]. This investigation caused great public concern [10].

The case of PFOA shows a long period of emergence of concern in the Netherlands, from the first articles reporting on the presence of PFOA in the environment in the early 2000s and the replacement of PFOA by another perfluorinated compound in 2012. Lau et al. [38] reviewed the literature on monitoring and toxicological findings about perfluoroalkyl acids in 2007. Based on this review, it can be concluded that Hansen et al. [39] quantitatively reported the presence of PFOA in the aquatic environment for the first time in 2002. However, we found that Moody et al. [40] had published research somewhat earlier in 2001, reporting the presence of PFOA in surface water samples. Another early paper on the presence of perfluorooctane surfactants in surface water, was the study by Boulanger et al. [41] who reported concentrations of PFOA in Great Lakes water.

The proposed methodology including the pattern shown in Additional file 1 was run for articles published between 2001 and 2007. The methodology did not pick up the articles by both Hansen et al. [39] (published in 2002) and Moody et al. [40] (published in 2001), because they did not specifically refer in either the title or the abstract to this being the first report of PFOA in the aquatic environment. However, the study by Giesy and Kannan [42] (published in 2001) on the presence of PFCs in (aquatic) wildlife was picked up by the proposed methodology. However, these authors focused primarily on providing evidence of the global distribution of perfluorooctane sulphonic acid (PFOS) in biota not so much a first reporting. Also, the article by Boulanger et al. [41] published 3 years later in 2004 was picked up. Thus, using the proposed text mining methodology, attention could have been drawn to the potential presence of PFOA in the aquatic environment in the Netherlands some 8 years earlier (in 2004 instead of 2012) and proactive risk governance at a national level would have been possible.

Legionella

Legionella bacteria are ubiquitously present in the environment. Inhaling pathogenic *Legionella* bacteria can

cause Legionnaires´ disease (LD) resulting in severe pneumonia. In 2017, the highest number of patients suffering from LD ever notified in the Netherlands was reported, namely a total of 561 cases [43], and only a minority of these were associated with exposure abroad. LD is often associated with manmade water systems, for instance, whirlpools, cooling towers and water distribution systems. However, the infection source remains unknown for most of the cases that are not part of an outbreak of Legionnaires´ disease and that have been infected in the Netherlands [43].

In 2016 and 2017, two successive clusters of a total of 14 cases of LD were reported in Boxtel, a town in the south of the Netherlands [44]. At first, no common source could be identified based on interviews and sampling. However, after continuously investigating possible sources, an industrial biological WWTP was identified as the infection source for both clusters. The growing trend in LD cases in another city in the south of the Netherlands was also traced back to an industrial biological WWTP. These findings illustrated the importance of industrial biological WWTPs as potentially relevant sources for LD infections [43].

In 2018, Loenenbach et al. [44] reported identifying industrial biological WWTPs as potential relevant sources of Legionnaires´ disease infections for the first time in the Netherlands. However, cases of Legionnaires´ disease with biological WWTPs as infection source had already been reported in other countries before the two successive clusters in the Netherlands in 2016 and 2017 were found. Indeed, van Heijnsbergen et al. [45] also mentioned these cases in their review of potential sources of Legionella which was published in 2015. To the best of our knowledge, Allestam et al. [46] identified the biological treatment of industrial wastewater as a possible source for *Legionella* infection for the first time in 2006.

The proposed methodology including the pattern shown in Additional file 1 was run for articles published between 2006 and 2015. The methodology did not pick up the research by Allestam et al. [46] (published in 2006), because it was not published as a scientific article, but as a book chapter. However, a Finnish report on two cases of Legionnaires´ disease associated with biological WWTPs published in 2010 [47] was identified. Thus, if the proposed text mining methodology had been used in the Netherlands, the potential significance of biological WWTPs in Legionnaires´ disease infection could have been identified in 2010 instead of 2015. In that case, the period of concern would have been decreased by 5 years and proactive risk governance would have been possible, for example, by running a monitoring campaign to identify relevant industrial biological WWTPs in the Netherlands.

Discussion

To the best of our knowledge, this is the first attempt to develop a methodology to search the scientific literature for articles reporting the first detection of chemical and microbial contaminants in the aquatic environment. Sjerps et al. [21] used text mining in 2015 to identify potential emerging risks, comparing the manual and automated analysis of scientific literature. The authors concluded that the manual analysis was not structured, poorly reproducible and labour-intensive. The automated search using the text mining tool was fast and reproducible but generated too many hits and an unmanageable number of contaminants. Therefore, Sjerps et al. [21] suggested using automated text analysis to identify eligible studies and then performing a manual analysis of the eligible studies. Using the pattern matching approach in this study is one way of implementing this as a reproducible methodology.

In this research project, we showed the results of applying the developed methodology to literature published in the last 2.5 years (2016 until August 2018). This resulted in 3650 records which were manually analysed using the additional predefined inclusion and exclusion criteria. Although the developed methodology minimised the manual workload as only sentences matching the pattern were analysed and not the whole abstract, this is still a time consuming step in the analysis. Therefore, in order to keep the number of records manageable, we suggest running the methodology twice a year. Based on the number of relevant articles published between 2016 and August 2018 (2016 = 157, 2017 = 137 and until August 2018 = 74), this would result in about 70 to 80 articles per run.

The effectiveness of the methodology was tested using a priori selected articles. One of the a priori selected articles, namely Conley et al. [27], was not found by the developed methodology. This is because the first detection of norfluoxetine was not mentioned in the abstract or title, but only in the full text. Therefore, by using the developed methodology only those articles are identified, in which the authors consider the first detection of a contaminant in the aquatic environment an important aspect of their research and include this in the title or abstract. Open Access publishing would remove this limitation as the full text could then be retrieved from Scopus® instead of the abstract (see code shown in Additional file 1). The added-value of text mining full text articles instead of abstracts has been illustrated before by Westergaard et al. [48]. However, a recent estimation of Open Access publishing showed that only 28 percent of scientific articles are published Open Access [49]. Thus, the limitation of mining only title and abstracts is not expected to be eliminated any time soon.

The specificity analysis resulted in a low specificity (52.1%). This is due to the high fraction of false positives. The calculation of the low specificity is once again evidence for the need of the additional manual analysis of the identified articles, as is shown in Fig. 1. Also, words are used in many different ways in a sentence, such as the words 'new' and 'first', which leads the pattern to extract false positives. For example, 'new' could be part of a region´s or city´s name, such as 'New Zealand' in the abstract published by Neary and Baillie [50]. The word 'first' is also used in many articles as a numerical transition word, for example in the abstract by Sharma and Malaviya [51]. Most false positives are unavoidable and can easily be excluded in the manual selection phase of eligible studies.

However, some of the false positives could be automatically eliminated by removing sentences in which "New" refers to a country and "first" is used in the beginning of a sentence and following by a comma. These rules were translated into additional lines of code (see Additional file 1) which could be run after the pattern matching code. We were able to automatically eliminate 161 sentences by using this additional line of code on the sentences shown in Additional file 3.

The fraction of false negatives found was very low, namely 0.0183. However, all false negatives reported on the first detection of a microbial contaminant indicating that the pattern is more tailored to studies reporting on chemical contaminants than to studies reporting on microorganisms in the aquatic environment. This can be due to the fact that the a priori selected articles comprised only two articles reporting on the first detection of microbial contaminants in the aquatic environment [52, 53]. Therefore, we suggest an addition to the pattern shown in Additional file 1, namely a combination of the words 'novel', 'new' or 'undescribed' and 'species', 'first outbreak' and 'first description'. The extended pattern is also available in Additional file 1 and eliminates 29 out of the 32 false negatives.

The methodology was made as straightforward as possible and coded in R to make it widely applicable. However, as the methodology is R-based, some prior knowledge of programming is needed to be able to run it. Therefore, we suggest researchers use the methodology to inform policy makers. For example, researchers working in close collaboration with national or international government agencies, such as employees of health agencies. Another option is to build a user interface as has been done previously for complicated computational analysis tools such as QMRAspot [54, 55]. These tools include data, assumptions and calculations which make them more user-friendly for non-mathematicians. However, it should be noted that, in order to interpret the results of these tools, discipline related knowledge is still required.

A retrospective validation of the methodology was performed by evaluating the period of emergence of concern for two example contaminants in the Netherlands, one microbial and one chemical contaminant. Although we are aware of the fact that the period of emergence of concern related to these contaminants might be very different in other countries and that early identification of contaminants is no guarantee for regulatory actions, the retrospective validation illustrated that the methodology can be useful for the more timely identification of emerging contaminants.

Although the methodology has been developed specifically to extract articles from Scopus®, any database of peer-reviewed literature could be used with the proposed search query. In that case, the developed code could be used as is after the abstract and title information has been imported into R-studio. However, to our knowledge, no R-package exists for retrieving abstract information from databases of peer-reviewed literature except for Scopus®.

Furthermore, the search query and pattern can be easily adjusted as the codes are added as supplementary material and the additional inclusion and exclusion criteria are explicitly described in Additional file 2. For instance, the search query and additional inclusion and exclusion criteria can be adjusted to make the methodology applicable to the search for articles identifying contaminants for the first time in soil or air. Identifying early signals of contaminants in soil might also be interesting when it comes to the quality of freshwater resources due to potential leaching. Also, by replacing all search terms in concept 1 of the search query (see Fig. 2) by a specific contaminant group, such as "pharmaceuticals" or "personal care products", the methodology could be used to identify a specific type of new chemicals. Finally, one might consider including studies on new toxicity results for known contaminants, and compare these to the results of national monitoring studies. In these cases, the pattern could be used as it is as long as the search terms are adapted.

When textual data were imported into the R environment, some characters were not properly encrypted and were thus replaced by random signs. Examples of characters that the R environment was unfamiliar with, even after an encryption comment was run, were Greek letters and characters in subscript or superscript. This phenomenon has caused some contaminants in the abstracts shown in Additional file 3 to be named incorrectly. However, as the Scopus® link to the original research is included in Additional file 3, the name of the contaminant can always be checked.

Finally, the developed methodology can be used to identify signals in any national, international or river basin setting since the search query and inclusion and

exclusion criteria are not country or area specific. However, it is recognised that the elucidation of the relevance of the signals in the national, international or river basin setting is a crucial part of the proactive governance of emerging contaminants in the aquatic environment. Only when the identified signals are analysed effectively, is proactive governance possible.

Conclusions

In this study, we hypothesised that the period of emergence of concern of contaminants could be reduced by performing a systematic search for articles which reported the first detection of a contaminant in the aquatic environment. For this purpose, we developed a methodology using literature mining. The technical aspects of the developed methodology were described as well as its implementation for the screening of recent scientific literature. The hypothesis was tested by retrospectively analysing the period of the emergence of concern related to two contaminants in the Netherlands. The retrospective analysis showed that the methodology is able to extract early signals of a contaminant in the aquatic environment. However, the further elucidation of the relevance of the identified signals, here referred to as the reporting phase, is crucial in order to decrease the period of emergence of future contaminants. We therefore conclude that the developed methodology is a first step towards the proactive systematic identification of emerging contaminants in the aquatic environment.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13750-019-0177-z.

Additional file 1. Scripts mentioned in manuscript Use of literature mining for early identification of emerging contaminants in freshwater resources by Hartmann et al. R-based codes.

Additional file 2. Overview of the search query and the inclusion and exclusion criteria. Table including search query, inclusion and exclusion criteria.

Additional file 3. Results of applying the developed methodology to recent literature, namely articles published between 2016 and 27th of August 2018. Table containing the articles' Electronic Identifier (EID), authors, title, publication year, journal, volume, page information, citations, Digital Object Identifier (DOI), link to the article in Scopus®, abstract, sentence matching the pattern, whether or not the article was considered to be eligible or not and why.

Additional file 4. Analysis for false negatives of random selection of 1750 articles from the 23,217 articles (published between 2006 and 2012) that did not match the pattern. Table containing the articles' Electronic Identifier (EID), authors, title, publication year, journal, volume, page information, citations, Digital Object Identifier (DOI), link to the article in Scopus®, abstract, the sentence that matched the pattern and whether the articles is considered to be a false negative or not.

Acknowledgements
We would like to thank all members of the project team of PS-DRINK for their valuable comments during the development process of the methodology. Also, we thank Els Smit, Saskia Rutjes and Theo Traas (RIVM) for their useful suggestions on an earlier draft of this manuscript.

Authors' contributions
JH developed the theory, performed the analysis and took the lead in writing the manuscript. SW, JPvdH, and AMdRH supervised the project. AMdRH helped to carry out the retrospective validation of the microbial contaminant. All the authors provided critical feedback and helped shape the research, analysis and manuscript. All authors read and approved the final manuscript.

Funding
This research was funded by the RIVM Strategic Research Project PS-DRINK (S/121014).

Availability of data and materials
The dataset(s) supporting the conclusions of this article are included within the article (and its additional file(s)).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] National Institute for Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, The Netherlands. [2] Delft University of Technology, PO Box 5, 2600 AA Delft, The Netherlands. [3] Copernicus Institute of Sustainable Development, P.O. Box 80125, 3508 TC Utrecht, The Netherlands. [4] Waternet, PO Box 94370, 1090 GJ Amsterdam, The Netherlands. [5] Institute for Risk Assessment Sciences, P.O. Box 80178, 3508 TD Utrecht, The Netherlands.

Received: 10 May 2019 Accepted: 28 September 2019
Published online: 13 October 2019

References

1. Gavrilescu M, Demnerova K, Aamand J, Agathos S, Fava F. Emerging pollutants in the environment: present and future challenges in biomonitoring, ecological risks and bioremediation. New Biotechnol. 2015;32(1):147–56.
2. Moritz S, Bunke D, Brack W, López Herráez D, Posthuma L. Developments in society and implications for emerging pollutants in the aquatic environment. Oeko-Institut Working Paper 1/2017; 2017.
3. Vouga M, Greub G. Emerging bacterial pathogens: the past and beyond. Clin Microbiol Infect. 2016;22(1):12–21.
4. Richardson SD, Kimura SY. Water Analysis: emerging Contaminants and Current Issues. Anal Chem. 2016;88(1):546–82.
5. Boxall AB, Rudd MA, Brooks BW, Caldwell DJ, Choi K, Hickmann S, et al. Pharmaceuticals and personal care products in the environment: what are the big questions? Environ Health Perspect. 2012;120(9):1221–9.
6. Bouki C, Venieri D, Diamadopoulos E. Detection and fate of antibiotic resistant bacteria in wastewater treatment plants: a review. Ecotoxicol Environ Saf. 2013;91:1–9.
7. Sabri NA, Schmitt H, Van der Zaan B, Gerritsen HW, Zuidema T, Rijnaarts HHM, et al. Prevalence of antibiotics and antibiotic resistance genes in a wastewater effluent-receiving river in the Netherlands. J Environ Chem Eng. 2018. https://doi.org/10.1016/j.jece.2018.03.004.
8. van Wezel AP, van den Hurk F, Sjerps RM, Meijers EM, Roex EW, ter Laak TL. Impact of industrial waste water treatment plants on Dutch surface waters and drinking water sources. Sci Total Environ. 2018;640:1489–99.
9. Duizer E, Rutjes S, de Roda-Husman AM, Schijven JJE. Risk assessment, risk management and risk-based monitoring following a reported accidental release of poliovirus in Belgium, September to November 2014. Eurosurveillance. 2016;21(11):30169.

10. Hartmann J, van der Aa M, Wuijts S, de Roda Husman AM, van der Hoek JP. Risk governance of potential emerging risks to drinking water quality: analysing current practices. Environ Sci Policy. 2018;84:97–104.
11. Halden RU. Epistemology of contaminants of emerging concern and literature meta-analysis. J Hazard Mater. 2015;282:2–9.
12. Blaak H, van den Berg HHJL, Docters van Leeuwen AE, Italiaander R, Schalk JAC, Rutjes SA, et al. Emerging pathogenen in oppervlaktewater. RIVM rapport 703719049. 2010:47.
13. Lodder WJ, Rutjes SA, Takumi K, de Roda Husman AM. Aichi virus in sewage and surface water, the Netherlands. Emerg Infect Dis. 2013;19(8):1222–30.
14. La Rosa G, Fratini M, della Libera S, Iaconelli M, Muscillo M. Emerging and potentially emerging viruses in water environments. Annali dell'Istituto superiore di sanità. 2012;48:397–406.
15. Reemtsma T, Berger U, Arp HPH, Gallard H, Knepper TP, Neumann M, et al. Mind the gap: persistent and mobile organic compounds - water contaminants that slip through. Environ Sci Technol. 2016;50:10308–15.
16. Paynter R, Bañez LL, Berliner E, Erinoff E, Lege-Matsuura J, Potter S, et al. EPC methods: an exploration of the use of text-mining software in systematic reviews 2016. https://www.ncbi.nlm.nih.gov/books/NBK362045/#methods.s1. Accessed 26 Mar 2018.
17. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(1):5.
18. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. Methods (San Diego, Calif). 2015;74:97–106.
19. Octaviano FR, Felizardo KR, Maldonado JC, Fabbri SCPF. Semi-automatic selection of primary studies in systematic literature reviews: is it reasonable? Empir Softw Eng. 2015;20(6):1898–917.
20. Van de Brug F, Luijckx NL, Cnossen H, Houben G. Early signals for emerging food safety risks: from past cases to future identification. Food Control. 2014;39:75–86.
21. Sjerps R, Puijker L, Brandt A, ter Laak T. BTO 2015.059 Signals of emerging compounds (2014–2015) (in Dutch). KWR; 2015.
22. Marshall C, Brereton P, editors. Tools to support systematic literature reviews in software engineering: a mapping study. Empirical software engineering and measurement, 2013 ACM/IEEE International Symposium on; 2013: IEEE.
23. Crisan A, Munzner T, Gardy JL. Adjutant: an R-based tool to support topic discovery for systematic and literature reviews. bioRxiv. 2018. https://doi.org/10.1093/bioinformatics/bty722.
24. Sarma G. Scientific literature text mining and the case for open access. J Open Eng. 2017. https://doi.org/10.7287/peerj.preprints.2566v2.
25. Elsevier. 2019. https://www.elsevier.com/solutions/scopus. Accessed 1 Aug 2018.
26. Muschelli J. Gathering Bibliometric Information from the Scopus API using rscopus. R Journal. 2018. http://works.bepress.com/john_muschelli/7/.
27. Conley JM, Symes SJ, Schorr MS, Richards SM. Spatial and temporal analysis of pharmaceutical concentrations in the upper Tennessee River basin. Chemosphere. 2008;73(8):1178–87.
28. Fabbri S, Hernandes E, di thommazo A, Belgamo A, Zamboni A, Silva C. Managing Literature reviews information through visualization. In: ICEIS 2012 - Proceedings of the 14th international conference on enterprise information systems, vol. 2; 2012. p. 36–45.
29. Feinerer I. Introduction to the tm package text mining in R. 2018. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf. Retrieved 1 Mar 2019.
30. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. quanteda: an R package for the quantitative analysis of textual data. J Open Source Softw. 2018;3(30):774.
31. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2(1):37–63.
32. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74.
33. Buck RC, Franklin J, Berger U, Conder JM, Cousins IT, DeVoogt P, et al. Perfluoroalkyl and polyfluoroalkyl substances in the environment: terminology, classification, and origins. Integr Environ Assess Manag. 2011;7(4):513–41.
34. Grandjean P, Clapp R. Changing interpretation of human health risks from perfluorinated compounds. Public Health Rep. 2014;129(6):482–5.
35. Calafat AM, Kuklenyik Z, Reidy JA, Caudill SP, Tully JS, Needham LL. Serum concentrations of 11 polyfluoroalkyl compounds in the U.S. population: data from the national health and nutrition examination survey (NHANES). Environ Sci Technol. 2007;41(7):2237–42.
36. Zeilmaker MJ, Janssen P, Versteegh A, van Pul A, de Vries W, Bokkers B, et al. RIVM report 2016-0049 Risk assessment of the emission of PFOA (in Dutch). 2016, p. 68.
37. Bokkers BGH, Versteegh JFM, Janssen PJCM, Zeilmaker MJ. Risk assessment of the exposure to PFOA via drinking water at two Locations (in Dutch). In: RIVM, editor. Letter 0150/2016/M&V/EvS/AV to the Ministery of Infrastructure and the Environment; 2016.
38. Lau C, Anitole K, Hodes C, Lai D, Pfahles-Hutchens A, Seed J. Perfluoroalkyl acids: a review of monitoring and toxicological findings. Toxicol Sci. 2007;99(2):366–94.
39. Hansen KJ, Johnson HO, Eldridge JS, Butenhoff JL, Dick LA. Quantitative characterization of trace levels of PFOS and PFOA in the Tennessee river. Environ Sci Technol. 2002;36(8):1681–5.
40. Moody CA, Kwan WC, Martin JW, Muir DCG, Mabury SA. Determination of perfluorinated surfactants in surface water samples by two independent analytical techniques: liquid chromatography/tandem mass spectrometry and 19F NMR. Anal Chem. 2001;73(10):2200–6.
41. Boulanger B, Vargo J, Schnoor JL, Hornbuckle KC. Detection of perfluorooctane surfactants in great lakes water. Environ Sci Technol. 2004;38(15):4064–70.
42. Giesy JP, Kannan K. Global distribution of perfluorooctane sulfonate in wildlife. Environ Sci Technol. 2001;35(7):1339–42.
43. Reukers D, van Asten L, Brandsema P, Dijkstra F, Donker G, van Gageldonk-Lafeber A, et al. Annual report Surveillance of influenza and other respiratory infections: Winter 2017/2018. 2018.
44. Loenenbach AD, Beulens C, Euser SM, van Leuken JP, Bom B, van der Hoek W, et al. Two community clusters of Legionnaires' disease directly linked to a biologic wastewater treatment plant, the Netherlands. Emerg Infect Dis. 2018;24(10):1914.
45. van Heijnsbergen E, Schalk JA, Euser SM, Brandsema PS, den Boer JW, de Roda Husman AM. Confirmed and potential sources of Legionella reviewed. Environ Sci Technol. 2015;49(8):4797–815.
46. Allestam G, de Jong B, Långmark J. Biological treatment of industrial wastewater: a possible source of Legionella infection. Legionella: American Society of Microbiology; 2006. p. 493–6.
47. Kusnetsov J, Neuvonen L-K, Korpio T, Uldum SA, Mentula S, Putus T, et al. Two Legionnaires' disease cases associated with industrial waste water treatment plants: a case report. BMC Infect Dis. 2010;10(1):343.
48. Westergaard D, Stærfeldt HH, Tønsberg C, Jensen LJ, Brunak SJPcb. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018;14(2):e1005962.
49. Pirmez C, Brandão AA, Momen H. Emerging infectious disease and fast-track publication: when public health gets priority over the formality of scholarly publishing. Mem Inst Oswaldo Cruz. 2016;111(5):285.
50. Neary DG, Baillie BR. Cumulative effects analysis of the water quality risk of herbicides used for site preparation in the Central North Island, New Zealand. Water. 2016;8(12):573.
51. Sharma S, Malaviya P. Bioremediation of tannery wastewater by chromium resistant novel fungal consortium. J Ecol Eng. 2016;91:419–25.
52. Su H-C, Ying G-G, Tao R, Zhang R-Q, Zhao J-L, Liu Y-S. Class 1 and 2 integrons, sul resistance genes and antibiotic resistance in Escherichia coli isolated from Dongjiang River, South China. Environ Pollut. 2012;169:42–9.
53. Hamza IA, Jurzik L, Wilhelm M, Überla K. Detection and quantification of human bocavirus in river water. J Gen Virol. 2009;90(11):2634–7.
54. Schijven JF, Teunis PF, Rutjes SA, Bouwknegt M, de Roda Husman AM. QMRAspot: a tool for quantitative microbial risk assessment from surface water to potable water. Water Res. 2011;45(17):5564–76.
55. Schijven J, Derx J, de Roda Husman AM, Blaschke AP, Farnleitner AHJJ. QMRAcatch: microbial quality simulation of water resources including infection risk assessment. J Environ Qual. 2015;44(5):1491–502.

Publisher's Note