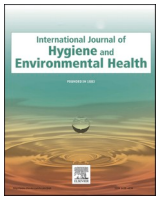
International Journal of Hygiene and Environmental Health 245 (2022) 114018

Contents lists available at ScienceDirect

International Journal of Hygiene and Environmental Health

journal homepage: www.elsevier.com/locate/ijheh



Using Artificial Intelligence to extract information on pathogen characteristics from scientific publications

Sotirios Paraskevopoulos [a,b,*], Patrick Smeets [a], Xin Tian [a], Gertjan Medema [a,b]

[a] KWR Water Research Institute, Groningenhaven 7, P.O. Box 1072, 3430 BB, Nieuwegein, the Netherlands
[b] Department of Water Management, Delft University of Technology, Stevinweg 1, 2628, CN Delft, the Netherlands

ARTICLE INFO

Keywords:
Artificial intelligence
Information extraction
Exposure assessment
Scientific publications
Legionella

ABSTRACT

Health risk assessment of environmental exposure to pathogens requires complete and up to date knowledge. With the rapid growth of scientific publications and the protocolization of literature reviews, an automated approach based on Artificial Intelligence (AI) techniques could help extract meaningful information from the literature and make literature reviews more efficient. The objective of this research was to determine whether it is feasible to extract both qualitative and quantitative information from scientific publications about the waterborne pathogen *Legionella* on PubMed, using Deep Learning and Natural Language Processing techniques. The model effectively extracted the qualitative and quantitative characteristics with high precision, recall and F-score of 0.91, 0.80, and 0.85 respectively. The AI extraction yielded results that were comparable to manual information extraction. Overall, AI could reliably extract both qualitative and quantitative information about *Legionella* from scientific literature. Our study paved the way for a better understanding of the information extraction processes and is a first step towards harnessing AI to collect meaningful information on pathogen characteristics from environmental microbiology publications.

1. Introduction

Human exposure to pathogens in the environment poses risks to public health (Hrudey and Hrudey, 2004). Health risk assessments are used to prevent or manage these risks and support decisions, for example on safe system design or emergency response. Exposure assessment is a first step in which knowledge about pathogen characteristics and their exposure routes are combined to estimate the exposure of the population to pathogens. With the fast-growing rate of scientific publications, such information is contained in a constantly increasing volume of text and journal articles. The conventional way is to generate review papers and meta-analyses to collate the published information, analyze the body of information in a comprehensive and integrated manner, and conduct such meta-analyses in an increasingly structured framework (Page et al., 2021). This process is time-consuming, labor-intensive and requires an expert that knows where to look and what to search for. The increasing rate of those publications has created a need for more efficient and extensive methods to collect all meaningful information for health risk assessment from various sources.

In recent years, automated approaches using Artificial Intelligence (AI) have been explored to systematically extract structured information from the ever-expanding body of scientific publications. Experts and curators in the field of biomedical sciences have been using AI and in particular Information Extraction (IE) techniques to extract information from Electronic Health Records (EHR) and Randomized Control Trials (RCT) (Cohen and Hersh, 2005; Meystre et al., 2008). Using text mining techniques (and consequently IE), Machine Learning (ML) and Natural Language Processing (NLP), experts extract information related to study characteristics such as disease-drug associations from EHR and RCT (Chen et al., 2008; Chung and Coiera, 2007; Kang et al., 2019; Uzuner et al., 2010). Kiritchenko et al. (2010), provided ExaCT, an IE system that extracts 21 key trial characteristics from publications and helps curators review and collect information from RCT (using a user interface). Their approach was based on ML using a Support Vector Machine (SVM) model for their sentence classification as well as rule-based techniques to extract exact values from segments within a text. A similar approach was adopted by Patrick and Li (2010), who used a multistage ML-based method with 2 different statistical classifiers namely SVM and Conditional Random Fields (CRF) and rule-based methods, they achieved an almost-optimal result (relative to other participants) for automated extraction of medication information from clinical notes. Although in the field of biomedical sciences, using such techniques (AI, IE, ML, and NLP) to extract information from text

* Corresponding author. KWR Water Research Institute, Groningenhaven 7, P.O. Box 1072, 3430 BB, Nieuwegein, the Netherlands.
E-mail address: Sotirios.Paraskevopoulos@kwrwater.nl (S. Paraskevopoulos).

https://doi.org/10.1016/j.ijheh.2022.114018
Received 22 February 2022; Received in revised form 29 July 2022; Accepted 30 July 2022
Available online 16 August 2022

Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CRF	Conditional Random Fields
DL	Deep Learning
EHR	Electronic Health Records
IE	Information Extraction
IK	Information Keywords
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomized Control Trials
Regex	Regular expressions
RNN	Recurrent Neural Networks
SVM	Support Vector Machine

documents has become a well-established approach; the development of similar applications in the field of environmental microbiology is still lagging and more complex because of the arbitrary and diverse form and structure in which the information is contained in case studies, reviews, and publications. The desired information is more scattered and complex compared to the structured information often contained in RCT and EHR. The less structured organization of the information requires an improved AI system that unravels the complexity of words and sentences by "understanding" and capturing the syntactic and semantic context of their surrounding words prior to the classification task.

This study aimed to evaluate the feasibility and performance of using an IE model to extract both qualitative and quantitative information about the waterborne pathogen *Legionella* from scientific publications. *Legionella* was selected since it is frequently associated with outbreaks via different water sources, many (types of) publications are available, and scientists and experts would like to have as much high quality information as possible to support decision making (van Heijnsbergen et al., 2015; Walser et al., 2014) and risk assessment (Papadakis et al., 2018).

To capture the information on *Legionella* as it is arbitrarily expressed in scientific literature, Deep Learning approach was developed in this study (instead of using the conventional classifiers used in ML), coupled with a rule-based technique. The quality of the extracted qualitative and quantitative information on *Legionella* was assessed using the evaluation metrics of precision, recall and F-score (Kiritchenko et al., 2010), along with a comparison between the system extraction and a human (manual) extraction.

2. Materials and method

2.1. Information keywords

The desired information (hereafter referred to as "information keywords") about *Legionella* was selected as general, explicit, and reproducible (waterborne) pathogen characteristics of both a qualitative and a quantitative nature (Table 1).

2.2. Selection of publications

50 peer-reviewed scientific publications about *Legionella* were manually selected from the search engine PubMed and used for the implementation of the IE task. We specifically aimed to extract information from peer-reviewed scientific publications, since this better warrants the quality of the text that we use for data extraction. The type of selected publications includes both scientific reviews and case studies on waterborne outbreaks, covering the different aspects of research on *Legionella*. A systematic review of the literature was performed adopting the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Liberati et al., 2009). The selection of publications was made considering their relevance to *Legionella* as well as their maximum possible reference to the desired Information Keywords (IK). The list of selected publications, the search terms, along with the flow diagram that describes the search process and the exclusion criteria can be found in the supplementary material.

Table 1
The desired extracted information (Information Keywords) from scientific publications regarding the waterborne pathogen Legionella. Incubation period is quantitative information whereas the rest information keywords are qualitative.

Information keywords	Description
Incubation period	The time elapsed between exposure to a pathogenic organism and symptom onset
Symptoms	The change in normal functions of a person indicating the presence of a disease
Clinical manifestations	The medical conditions of a patient after infection by the pathogen
Sources of exposure	Places or objects that spread the pathogen
Route of transmission	Route via which an individual became exposed to the pathogen
Environmental habitat	The environment/water system in which the pathogen grows
Species	Unit of classification and taxonomic rank of an organism

2.3. Template filling of the information

Template filling is an efficient approach (especially when the content of a text document describes an event or a situation) to extract information in a comprehensive, structured form. The process of template filling includes identifying and locating predefined entities and filling in their template slots. Table 2 depicts an example of template filling. The algorithm behind the template filling should be able to fill in the slots for both qualitative and quantitative information. However, not every slot can always be filled since it is possible that some IK might not be addressed in the text document. The IK vary in terms of their structure. Some consist of straightforward information such as "incubation period", and others, such as "Route of transmission" or "Environmental habitat" consist of lengthy, more vague, and free text information.

2.4. Information extraction task

2.4.1. Labeling and training the data

The first step of the IE task was to manually label the scientific publications. The labeling of data is part of the custom-trained NER model that requires a token-level classification, and it helps assess

Table 2
Example of template filling extracting information from a scientific publication.

Information keywords	Results
Species	*Legionella pneumophila*
Incubation period (days)	2–14
Symptoms	Headache, myalgia, asthenia, anorexia, fever, cough, chills, dyspnea, arthralgia
Route of transmission	Inhalation, micro aspiration, direct contact with surgical wounds
Environmental habitat	Aquatic habitats, water distribution systems
Clinical manifestation	Legionnaires' disease, atypical pneumonia, Pontiac fever
Source of exposure	Water supply, infectious aerosols, cooling towers, hot tubs, potting soil

whether a specific world within a sentence is relevant to a specific IK. Relevant words are those who are assigned to one of the IK labels, whereas irrelevant tokens are those who have no meaning to the labeling process and are assigned the label "O".[1] Fig. 1 serves as an example of the labeling process.

Next, the training and classification of labeled data was necessary so that the system will learn to correctly assign the right labels to words within sentences. This step was implemented using Python programming language (Van Rossum and Drake Jr, 1995) and the Spacy library (Honnibal, M., & Montani, 2017). The selection of Spacy library was made mainly because this tool is suitable for NLP tasks utilizing word embedding methods as well as Recurrent Neural Networks (RNN) for multiclass classification.

2.5. Overall architecture

Fitting the overall architecture into a general workflow resulted in the following process (Fig. 2).

2.5.1. Text pre-processing

Although scientific publications come in various document standards and formats, the 50 selected scientific publications were extracted from the PubMed search engine in a PDF format. The first step of the pre-processing process was the conversion of PDF files to text files so that they can be recognized and processed as raw data. Next, all the sections from the text documents that are irrelevant to the IE task were removed automatically. That includes references, editors' notes, and acknowledgments. It was decided that the summary of publications should also be excluded since the contained information can be found in the remaining sections of the text. To detect these sections ("References", "Acknowledgements", and "Summary") we assumed a consistency in the way the headings were expressed in the scientific publications before applying a rule-based keyword matching technique to filter them out. The cleaning process also included the conversion of all uppercase letters to lowercase, and removal of punctuation. The last step was the tokenization of words to facilitate the labeling process as well as the implementation of the model itself.

2.5.2. Rule-based techniques

For the IK "incubation period", regex pattern-matching was selected using a specific module embedded in Python (Kuchling, 2002). The information is in numeric form and follows a certain pattern in the text (e.g. "*the incubation period was 2 to 14 days", "the incubation ranges between 2 to 14 days prior to symptom onset*"). After isolating the sentences containing the word "incubation" from the text, a set of regular expressions was applied to every sentence for the extraction of digits or a range of digits that correspond to the number of days of the incubation period. For IK "symptoms" and "species", a pool parsing technique was adopted. Since the results of these 2 IK are finite and known, a pool with all the potential symptoms and species associated with *Legionella* was created. Then, during parsing of unseen text, several n-grams were matched each time to the pools to determine if any of the potential symptoms and species of the pool can also be found in the text document of interest. For the creation of the symptoms and species pool, all the potential symptoms and species (both pathogenic and non-pathogenic) associated with *Legionella* and Legionnaire's disease were collected after exploring the literature.

2.5.3. Supervised technique

For the remaining of IK, a supervised technique was used since the information to be extracted was neither confined within a finite set nor could be represented in a certain pattern of strings (as in the case of IK "incubation period", "symptoms", and "species"). The extraction of such information was therefore only possible by understanding the semantic pattern and relationship of the tokens[2] within a text document. Specifically, a custom-trained NER model using word embedding and RNNs was implemented. During the training process, after embedding the tokens (words) into a sequence of vectors (numerical representation of text), bidirectional RNNs were used to take the semantic context into consideration by encoding the vectors into a context-sensitive sentence matrix. Next, to improve the power of the model the system used an attention mechanism where the previously produced matrix was reduced to a sentence vector by selecting the most "appropriate" information (after applying weights to every token based on their importance). In the last step, after all text was converted to a single vector, the system was able to predict the classes of every token. This four-step formula named: "Embed, encode, attend, predict" is the fundamental approach adopted in Spacy library for NER and more documentation can be found in Honnibal (2016).

2.5.4. Post-processing of results

After the supervised and rule-based techniques had completed their task, the extracted information filled the slots of a pre-defined template comprised of the desired IK. The extracted information might consist of repeated words or words that have the same semantic meaning but differ in the length of characters in the text. For example, the slot of IK "Clinical Manifestation" may have both "Legionnaires Disease" and "Legionnaire's disease" in the template. Although the semantic meaning is the same, the two extracted sequences differ slightly (apostrophe). Therefore, to avoid extracting duplicate information, we used the Levenshtein distance, a string metric that measures the pattern similarity -or to put it differently- the differences between words and/or sequences of words (Levenshtein, 1966). Using the Levenshtein Python C extension module, the system decided whether or not to keep the extracted similar words in the template (Necas, D., Ohtamaa, M., Haapala, 2014).

2.6. Evaluation of the performance

The last step was the evaluation of the model output. To get an unbiased performance of the model, a 5-fold cross-validation method was implemented. After the system was trained by feeding it with 40 text documents (80% of total publications), the NER model was tested by using a set of 10 "unseen" testing data (20% of total publications). This process was repeated 5 times, each time with a separate set of training and testing data. For every iteration, the manually labeled values were compared with the predicted values for every IK in a so-called confusion matrix. Next, the evaluation metrics of precision, recall, and F-score were calculated to describe the performance of the model for that particular fold of data, and the metrics of all the folds were averaged to get the overall performance of the model.

The analytic approach of precision, recall, and F-score was adopted (Kiritchenko et al., 2010) and it was applied both to the system and to every IK separately after averaging the values through every fold (5 iterations). When it comes to classification tasks, precision is a metric that quantifies the number of correct positive predictions from all returned positive predictions. It is therefore the number of true positives divided by the number of true positives plus false positives (Equation (1)).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall, on the other hand, is a metric that quantifies the number of correct positive predictions made of all positive predictions that could have been made by the system. Specifically, it is the number of true

[1] The choice of the word "O" is a default option and it means that all the words irrelevant to the IK are automatically assigned to the label "O".

[2] Tokenization: In a sequence of characters within a text document, tokenization is the process of chopping up the sequence into pieces (words), named tokens (Webster and Kit, 1992).

TOKENS	Legionella	spp	are	ubiquitous	in	aquatic	habitats	and	water	distribution	systems.	The	symptoms	of	LD	are	fever	cough	and	chills.
LABELS	O	O	O	O	O	ENV. HABITAT	ENV. HABITAT	O	ENV. HABITAT	ENV. HABITAT	ENV. HABITAT	O	O	O	CLIN. MANIFEST	O	SYMPTOMS	SYMPTOMS	O	SYMPTOMS

Fig. 1. Example of the labeling process. The labels "Env. Habitat", "Clin. Manifestation", and "Symptoms" are assigned to their respective words, whereas the remaining irrelevant words have been assigned to the label "O".

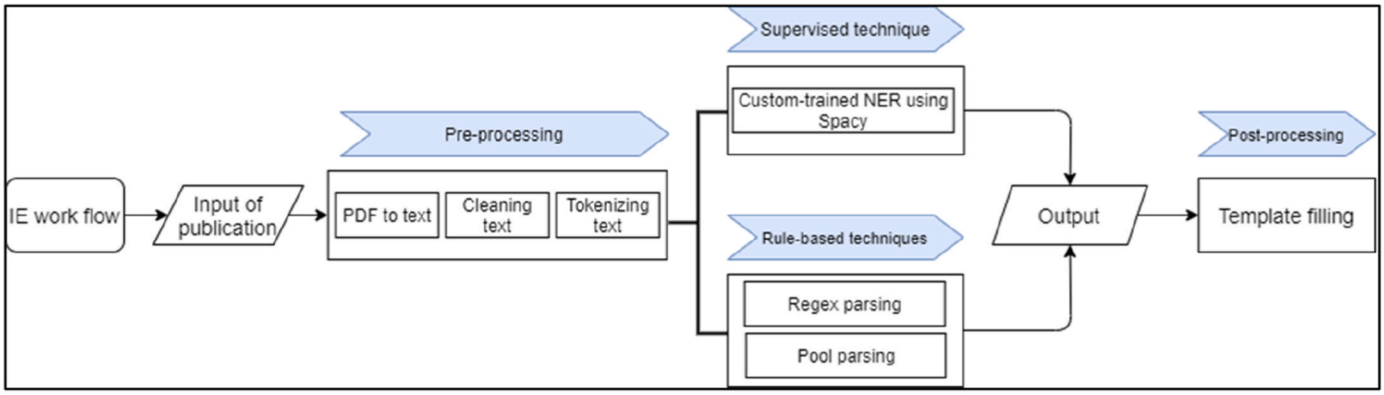


Fig. 2. The workflow of the IE task starts with the input of publication. Next, the publication gets converted to text, cleaned, and tokenized as part of the pre-processing step. The next part includes the supervised and rule-based techniques for the extraction of information. Finally, the output of this process gets filled in a template as part of the post-processing step. More can be found in chapters 2.5.1 -2.5.4.

positives divided by the number of true positives plus false negatives (Equation (2)).

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The F-score (Equation (3)) is the harmonic mean of precision and recall. It is a way to combine both analytic metrics into a single score that captures both properties (Olson and Delen, 2008).

$$F = 2^{*}\frac{Precision^{*}Recall}{Precision + Recall} \tag{3}$$

Choosing the right number of scientific publications for the training of the model was an important decision to make. Usually, the amount of data required to build a good DL model depends on the complexity of the problem (in our case extracting words and excerpts of information from unstructured scientific publications) and the quality of the training data. Regarding DL, the hypothesis is that the more quality data used to train a model, the higher is the performance (Mitsa, 2019). The impact of the number of publications used for training the IE model on the quality of the results was investigated. We created 5 folders containing 10, 20, 30, 40, and 50 publications randomly selected from the 50 papers that had been selected previously and performed a 5-fold Cross-validation in every folder.

Another form of evaluation was to select new publications (beyond the 50 that were used before) and compare the system's performance on IE with a manual extraction process (the conventional way where a human extracts information from text documents). We selected a set of 10 new scientific publications related to *Legionella* and incorporated them in the IE module. The same publications were processed by a human expert for manual extraction of the IK and the results were compared to assess the usefulness of the proposed approach on extracting information from *Legionella* scientific publications.

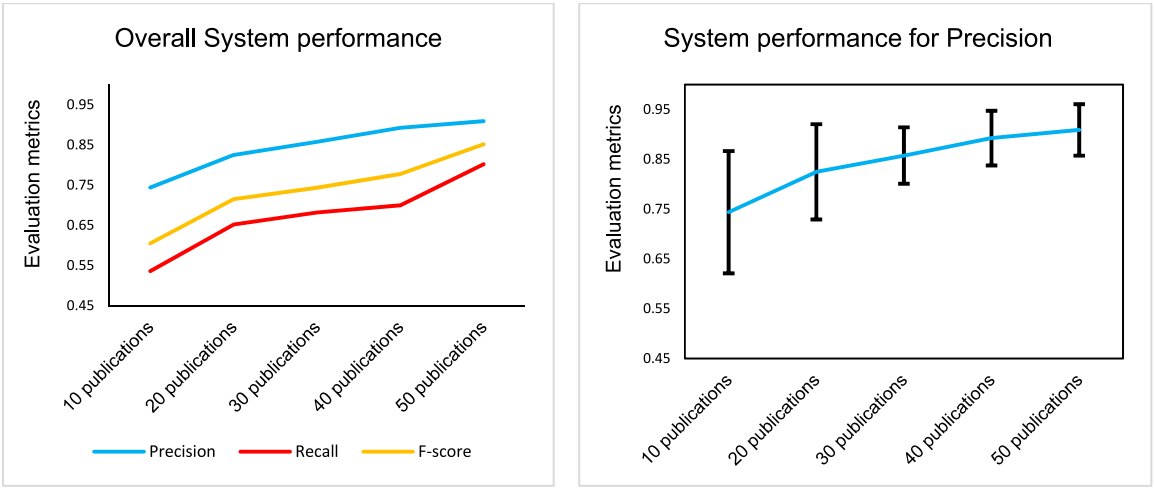


Fig. 3. a) System performance under different number of publications. b) System performance and standard deviation for precision under different number of publications.

3. Results

3.1. Influence of the number of publications on evaluation metrics

Fig. 3a shows that by increasing the number of publications, all metrics improved and the standard deviation of cross-validation regarding precision in Fig. 3b decreased overall (the standard deviation for recall and F-score can be found in the supplementary material). That means that by increasing the number of training data (publications) the model generalizes and thus, there is a smaller variation in its performance. These 2 interpretations go in line with the original hypothesis and since the standard deviation of precision remained constant for 3 consecutive increments of publications, we decided that 50 publications were an adequate and feasible starting point for the creation of the model. All further results were generated using 50 publications to train the model.

3.2. Evaluation of the supervised and rule-based extraction

For the supervised technique with custom-trained NER, the information on "Clinical manifestations", "Environmental habitat", "Route of transmission" and "Source of exposure" was extracted from the 50 publications. After performing a 5-fold cross validation to test the model, Table 3 shows the results of the 1st folder in a confusion matrix. The confusion matrix compares the actual with the predicted IK labels, indicating that the custom-trained NER technique was able to correctly predict the labels in the majority of the tokens. The only label that seemed to have mislabeled many features was the label "O" (which contains all the irrelevant words in a document). That "confusion" was expected to a certain extent since there was an imbalance between the label "O" and the rest of the IK (15897 tokens assigned to label "O" versus 2404 assigned to the rest of the IK) in the testing data. Considering that the desired information was generally organized in a complex and sparse manner within the text, it was expected to see false negatives. The label "O" affected and captured some of the words that should have been assigned to other labels. Another set of IK mislabeling their tokens were the "Source of exposure" and "Environmental habitat". This "confusion" was also expected since in many scientific publications the meaning of these two IK was often mixed and misinterpreted (i.e. "*The source of exposure of Legionella was 2 cooling towers*", " *Legionella* can grow and survive in cooling towers"). We see in this example that cooling towers can be labeled both as "Source of exposure" and "Environmental habitat" and therefore it was difficult for the system to always make correct predictions.

For the extraction of the information on "Incubation period", "Species", and "Symptoms" with rule-based techniques, almost all of the tokens were correctly labeled to their respective IK (Table 4). One IK that mislabeled some tokens, resulting in false negative results, was the "Incubation period". Looking into the testing dataset, this happened because in some publications, although the authors were describing the incubation period, they did not mention specifically the word "incubation" and therefore the regex rules did not apply. Another IK that mislabeled some tokens was the "Symptoms". Out of 521 tokens describing symptoms, 20 of them were not assigned correctly, probably because during the pool parsing technique, the respective pool did not contain those specific symptoms.

Table 4
Confusion matrix of the rule-based techniques.

	Predicted labels				
Actual labels	**Information keywords**	Incubation period	O	Species	Symptoms
	Incubation period	70	20	0	0
	O	0	46226	0	2
	Species	0	1	1011	0
	Symptoms	0	20	0	501

The classification reports in Tables 5 and 6, give an overview of the evaluation metrics of the system for the supervised and rule-based techniques. For the custom-trained NER in Table 5, the overall score of the system has a precision, recall, and F-score of 0.91, 0.80, and 0.85 respectively. While the precision score is high for IE tasks, the recall score of 0.80 leaves room for improvement (Patrick and Li, 2010; Kiritchenko et al., 2010). As explained earlier, the label "O" influenced to a certain extent the recall score of all individual IK (too many False Negatives for all IK), which resulted in a low overall score. The IK with the lowest metrics (both precision and recall) is the "Environmental habitat". This is because sometimes the environmental habitat of *Legionella* can also be presented as its source of exposure and vice versa. For the remaining IK, both precision and recall scores are high numbers.

For the rule-based techniques, as it was expected, the evaluation metrics for all IK are high with an overall precision and recall of 1 and 0.91 respectively.

3.3. Alternative evaluation with new publications

3.3.1. Improving the regex rules

After comparing the IE results with the human extraction, we identified a few setbacks on the proposed rule-based technique. Specifically, during the extraction of IK "Incubation period", the system could not distinguish the semantic difference between the actual incubation period of *Legionella* in patients prior to symptom onset, and the number of days required for the growth of colonies on solid media in a laboratory environment (a scientific publication can include both, i.e. *"L. gormanii and L. wadsworthii isolates resulted in no visible growth after 96 h incubation in BYE broth"*). Although both instances describe incubation period, their semantic is different. Therefore, a new set of rules was added that

Table 5
Classification report of the system's performance for the custom-trained NER.

Classification report	Precision	Recall	F-score	Total number of actual labels
Clinical Manifestation	0.95	0.88	0.91	725
Environmental habitat	0.81	0.73	0.77	286
Route of transmission	0.97	0.81	0.88	114
Source of exposure	0.91	0.79	0.85	1279
Average	0.91	0.80	0.85	–

Table 3
Confusion matrix of the custom-trained NER performance.

	Predicted labels					
Actual labels	**Information keywords**	Clin. Man/on	Env. habitat	O	Route of transmission	Source of Exposure
	Clin. Man/on	637	0	88	0	0
	Env. habitat	3	207	58	0	9
	O	32	31	15517	2	91
	Route of transmission	1	0	20	92	1
	Source of Exposure	2	19	273	1	984

Table 6
Classification report of the system's performance for the rule-based techniques.

Classification report	Precision	Recall	F-score	Total number of actual labels
Incubation period	1	0.78	0.88	90
Species	1	1	1	1012
Symptoms	1	0.96	0.98	521
Average	1	0.91	0.95	–

would exclude all mentions of *Legionella* associated with laboratory results.

3.3.2. Comparing the system with a human extraction

The alternative evaluation of the model (input of 10 new publications into the model and comparison with a human extraction) shows that the model returned results similar to the human extraction and extracted most of the IK from the text document. The classification report in Table 7 supports this argument. Although the sample is small and conclusions cannot be drawn, the evaluation metrics of both precision and recall are high. Table 8 depicts the extraction of information (and comparison) for 2 publications as example. The rest of the comparison tables can be found in the supplementary material.

4. Discussion

4.1. Evaluation of the IE model

The proposed IE model demonstrated very good performance on a set of 7 information keywords and extracted both quantitative and qualitative information regardless of the complexity of the targeted information. After testing it with 50 testing publications (10 publications per 5 folds of cross-validation) from various aspects of research on *Legionella* (scientific reviews and outbreak reports) the system was able to extract meaningful information. For the set of IK, both supervised and rule-based techniques were needed. The results of the evaluation metrics showed that the IE approach can adequately extract the desired information from scientific publications regarding the waterborne pathogen *Legionella*. Overall, the IE system identified and extracted the targeted IK with high precision (0.91) and provides proof of concept for automated extraction of this type of information from scientific publications. The lower recall score (0.80) indicated that the IE model missed some of the information. While the system's performance was not perfect and there is room for improvement, it is comparable with other IE tasks from biomedical sciences. In Kiritchenko et al. (2010), the results of precision and recall were 0.93 and 0.91 respectively whereas in Patrick & Li (2010), their precision had a score of 0.89 and recall 0.82. Finally, although not focused on NER, an IE task from tables in biomedical literature had 0.94 score for both precision and recall (Milosevic et al., 2019).

The alternative evaluation of the IE model confirmed the validity of our approach: when comparing the system's results with the manual extraction in 10 new publications on *Legionella*, the IE system returned similar results for all 7 IK. Although in some cases the IE model extracted irrelevant information for some of the IK, considering the complexity of the desired information, the results of the proposed IE model were of sufficiently high quality.

Table 7
Classification report of the custom-trained NER on the 10 new publications.

Classification report	Precision	Recall	F-score
Clinical Manifestation	0.76	0.91	0.81
Environmental habitat	0.63	0.92	0.71
Route of transmission	0.66	0.89	0.72
Source of exposure	0.68	0.87	0.75
Incubation period	1	0.75	0.83
Species	1	1	1
Symptoms	1	0.72	0.82
Average	0.82	0.87	0.81

4.2. Limitations and recommendations

Although the proposed approach showed promising results, it is accompanied by limitations. The main limitation stems from the very nature of the study's objective. IE tasks have not been implemented for data extraction on waterborne pathogens from scientific publications before. Therefore, there is still no relevant work to allow for a comprehensive comparison with the results of the proposed IE model. Although the proposed approach is based on similar work applied to biomedical data extraction using ML approaches, an established open-access benchmark dataset related to waterborne pathogens data extraction utilizing DL methods is missing. Considering the plethora of methods available in the literature for AI-data extraction using ML and DL methods, it is recommended that other approaches should also be tested.

Considering the proposed approach, the complexity of some of the IK is another limitation which resulted in missing some of the information (lower recall score). It was relatively easy to extract straightforward information, but when the desired information was unstructured, lengthy, or vague, the system sometimes failed to correctly identify its label. For example, for the IK "Clinical manifestation", the system would potentially have to target and extract words such as "Legionnaires' disease", "Pontiac Fever", and "pneumonia". The problem, in this case, is that the targeted fragment of words can be mentioned anywhere in a text document, each time in a different semantic context. Another limitation was the choice of pool parsing technique for the IK "Symptoms". Although the pool of symptoms included a variety of symptoms (more than 40), it was limited only to the symptoms collected manually from the literature. That means that there could be symptoms that the IE model would fail to recognize simply because they were not included in the respective pool. To tackle this limitation, an enrichment of the symptoms pool is recommended by incorporating all symptoms listed in the National Library of Medicine's Unified Medical Language System (UMLS) associated with the waterborne pathogen *Legionella* (Bodenreider, 2004). Finally, although the choice of regex rules showed good results, it also presented some difficulties in the information extraction process. The inability of the IE model to extract the incubation period in sentences where the word "incubation" is not mentioned, indicated the need for a slightly different approach. Instead of first isolating the word "incubation" from the whole text prior to applying the regex rules, it is recommended to first perform a sentence-level classification, extracting the sentences that contain the relevant information, and then apply the regex rules in the sentences that have been classified correctly. Doing that can ensure that all the values of the IK "Incubation period" can be extracted from the text.

4.3. Potential applications of IE tasks

Experts can use the IE model to extract high quality information in substantially less time (compared to the conventional way) for meta-analysis purposes. A meta-analysis can help recognize patterns, enrich the knowledge on *Legionella* (or other pathogens), and/or generate hypotheses. For example, by gathering information from multiple scientific publications (reviews and/or outbreak reports) regarding the incubation period of *Legionella*, it would be possible to create a distribution curve of the incubation time. Other examples are to collect and categorize various transmission pathways, or to identify the most common symptoms based on their frequency in Legionella outbreaks. Finally, by measuring the frequency of reported Legionellosis (the clinical manifestation of Legionella infection) case studies associated with exposure events, it is possible to estimate the likelihood of sources of exposure. All of these meta-analysis examples demonstrate the potential and

Table 8
Comparison between the system's performance and manual extraction of IK from 2 publications (Beauté et al., 2020; Couturier et al., 2020). Red highlighted shade = erroneous results. Red bold font = Missed result (either by the IE model or by the manual extraction).

	Healthcare-Associated Legionnaires Disease, Europe, 2008–2017. (Beauté et al., 2020)		*Transmission of Legionnaires Disease through Toilet Flushing* (Couturier et al., 2020)	
	System	Manual extraction	System	Manual extraction
Clinical manifestation	"ld" **"community acquired ld** "pneumonia" "legionnaires disease" **"knoxville"**	"Legionnaires disease" "pneumonia""ld"	"legionnaires disease" "pneumonia" "renal failure "bilateral pneumonia"	"legionnaires disease" "pneumonia" **"respiratory** and renal failure "bilateral pneumonia"
Environmental habitat	"human made water systems" "aquatic environments"	"aquatic environments" "human-made water systems"	**"hot** water" **"hospital hematology" "respiratory"**	"water"
Route of transmission	"inhalation" "aspiration"	"inhalation" "aspiration"	-	**"inhales" "person to person"**
Source of exposure	"potable water" "pools" **"nursing homes" "peaking"** "humidifiers" "decorative fountains" **"school" "activities"** **"composts"** **"demographic variables" "medical epidemiologist"** "potting soil" "medical devices"	"potable water" **"bathing" "steam-heated towels"** "humidifiers" "decorative fountains" "medical devices" "birthing pools"	"windshield washer fluid" "fountains" "dental unit" "cooling towers faucets" **"hospital" "aerosols" "contaminated** toilet" "air filtration" "shower" **"filters"**	"showers" "cooling towers" "faucets" "fountains" "windshield washer fluid" "dental unit" "waterlines" **"flushing toilets"** "air filtration systems" **"sink"** "toilet **water"**
Species	"L wadsworthii""anisa""L longbeachae" "L feeleii" "sainthelensi""micdadei" "pneumophila" "L bozemanii" "L dumoffii" "cincinnatiensis" "macechernii"	"L wadsworthii""anisa""L longbeachae" "L feeleii" "sainthelensi""micdadei" "pneumophila" "L bozemanii" "L dumoffii" "cincinnatiensis" "macechernii"	"L pneumophila"	"Legionella pneumophila"
Symptoms	-	-	"dyspnea" "fever"	"dyspnea" "fever" **"shivering"**
Incubation	"20 days"	**"2-10 days"** "20 days"	-	-

importance of using AI and specifically IE tasks to automatically extract high-quality information from scientific publications.

4.4. Future research

Future research should focus on improving the overall performance of the proposed approach. A hybrid system (a combination of the proposed DL method with another discriminative classifier such as CRF or SVM) could potentially improve the system's overall performance as previous research has shown (Lê, T., & Burtsev, 2019; Patrick and Li, 2010). For example, assigning the NER task to the custom-trained NER developed here and then coupling it with another classifier to classify relationships between entities could potentially further unravel the complexity of some of the IK. Another approach would be to consider using another DL approach, namely the Bidirectional Encoder Representations from Transformers (BERT). Based on the so-called Transformer neural network, this technique has gained attention and has become a ubiquitous baseline in NLP tasks, since it examines the context of words in both directions within a sentence (Kalyan et al., 2021).

4.4.1. Extrapolate the process to other pathogens and/or fields

Although this paper is focused on the waterborne pathogen *Legionella*, the IK are generic for waterborne pathogens. The good results with *Legionella* indicate that the IE model could also be successful for other waterborne pathogens, although many of those are not uniquely waterborne, but also spread via other matrices (food) or via person-to-person contacts, adding more complexity. The ability of DL methods (coupled with rule-based techniques) to unravel the complexity of information found in scientific publications enables experts to create more custom-train NER models using sufficient and representative training data from other waterborne pathogens publications. The proposed approach also enables scientists from different scientific domains to explore the power of using AI to extract complex, qualitative, or quantitative information from scientific publications. For example, the use of IE could be tested for the ability to extract functions such as inactivation rates (at different temperatures), disinfection kinetics, or log removal values of pathogens from various treatment processes found in scientific case studies.

5. Conclusions

This paper aimed to evaluate the feasibility and performance of a newly developed IE model to extract both qualitative and quantitative information from scientific publications about the waterborne pathogen *Legionella*. For the IE model, we adopted a combination of supervised (custom-trained NER model) and rule-based (regex pattern-matching, and pool parsing) techniques. The evaluation metrics showed a satisfactory performance for extraction of both qualitative and quantitative information: the custom-trained NER model had an overall F-score of 0.85, and the rule-based techniques had an F-score of 0.95. The IE model returned similar results with the manual extraction indicating that the extracted information is of high quality, and it can be further used by experts who seek to extract meaningful information from scientific publications using AI.

Overall, this study indicates that IE can provide an efficient and adequate approach for extracting qualitative and quantitative information on waterborne pathogen characteristics from the complex body of environmental microbiology literature. Scientists and experts can therefore begin to harness the power of Artificial Intelligence and Deep Learning techniques in this science field.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883484 PathoCERT Project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijheh.2022.114018.

References

Beauté, J., Plachouras, D., Sandin, S., Giesecke, J., Sparén, P., 2020. Healthcare-associated legionnaires' disease, Europe, 2008−2017. Emerg. Infect. Dis. 26, 2309. https://doi.org/10.3201/EID2610.181889.

Bodenreider, O., 2004. The unified Medical Language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, D267–D270. https://doi.org/10.1093/NAR/GKH061.

Chen, E.S., Hripcsak, G., Xu, H., Markatou, M., Friedman, C., 2008. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. J. Am. Med. Inf. Assoc. 15, 87–98. https://doi.org/10.1197/JAMIA.M2401/2/JAMIAM2401.F04.JPEG.

Chung, G.Y., Coiera, E., 2007. A study of structured clinical abstracts and the semantic classification of sentences. Biol. Transl. Clin. Lang. Process. 121–128.

Cohen, A.M., Hersh, W.R., 2005. A survey of current work in biomedical text mining Briefings Bioinf. 6, 57–71.

Couturier, J., Ginevra, C., Nesa, D., Adam, M., Gouot, C., Descours, G., Campèse, C., Battipaglia, G., Brissot, E., Beraud, L., Ranc, A.G., Jarraud, S., Barbut, F., 2020. Transmission of legionnaires' disease through toilet flushing. Emerg. Infect. Dis. 26, 1526. https://doi.org/10.3201/EID2607.190941.

Honnibal, M., Montani, I., 2017. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing | Sentometrics Research [WWW Document]. URL. https://sentometrics-research.com/publication/72/. accessed 12.17.21.

Honnibal, M., 2016. Embed, encode, attend, predict: the new deep learning formula for state-of-the-art NLP models · Explosion [WWW Document]. Explosion. URL. https://explosion.ai/blog/deep-learning-formula-nlp. accessed 12.17.21.

Hrudey, S.E., Hrudey, E.J., 2004. Safe Drinking Water. IWA Publishing, London, UK.

Kalyan, K.S., Rajasekharan, A., Sangeetha, S., 2021. AMMUS : A Survey of Transformer-Based Pretrained Models in Natural Language Processing.

Kang, T., Zou, S., Weng, C., 2019. Pretraining to recognize PICO elements from randomized controlled trial literature. Stud. Health Technol. Inf. 264, 188–192. https://doi.org/10.3233/SHTI190209.

Kiritchenko, S., De Bruijn, B., Carini, S., Martin, J., Sim, I., 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med. Inf. Decis. Making 10. https://doi.org/10.1186/1472-6947-10-56.

Kuchling, A.M., 2002. Regular expression HOWTO release 0.03 [WWW Document]. URL. http://www.python.org/doc/howto/. accessed 12.20.21.

Lê, T., Burtsev, M.S., 2019. A deep neural network model for the task of named entity recognition. Int. J. Mach. Learn. Comput. 9, 2019.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. 10, 707–710.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J. Clin. Epidemiol. 62, e1–e34. https://doi.org/10.1016/j.jclinepi.2009.06.006.

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb. Med. Inform. 17, 128–144. https://doi.org/10.1055/s-0038-1638592.

Milosevic, N., Gregson, C., Hernandez, R., Nenadic, G., 2019. A framework for information extraction from tables in biomedical literature. Int. J. Doc. Anal. Recogn. 22, 55–78. https://doi.org/10.1007/s10032-019-00317-0.

Mitsa, T., 2019. How do you know you have enough training data? [WWW Document]. URL. https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee. accessed 12.20.21.

Necas, D., Ohtamaa, M., Haapala, A., 2014. Levenshtein python c extension module [WWW Document]. URL. https://pypi.org/project/python-Levenshtein/. accessed 12.20.21.

Olson, D.L., Delen, D., 2008. Advanced Data Mining Techniques. Springer Science & Business Media.

Page, M.J., Moher, D., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., McKenzie, J.E., 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372, n160. https://doi.org/10.1136/bmj.n160.

Papadakis, A., Chochlakis, D., Sandalakis, V., Keramarou, M., Tselentis, Y., Psaroulaki, A., 2018. Legionella spp. risk assessment in recreational and garden areas of hotels. Int. J. Environ. Res. Publ. Health 15 (598 15), 598. https://doi.org/10.3390/IJERPH15040598, 2018.

Patrick, J., Li, M., 2010. High accuracy information extraction of medication information from clinical notes : 2009 i2b2 medication extraction challenge. https://doi.org/10.1136/jamia.2010.003939, 524-527.

Uzuner, Ö., Solti, I., Cadag, E., 2010. Extracting medication information from clinical text. J. Am. Med. Inf. Assoc. 17, 514–518. https://doi.org/10.1136/jamia.2010.003947.

van Heijnsbergen, E., Schalk, J.A.C., Euser, S.M., Brandsema, P.S., den Boer, J.W., de Roda Husman, A.M., 2015. Confirmed and potential sources of Legionella reviewed. Environ. Sci. Technol. 49, 4797–4815. https://doi.org/10.1021/acs.est.5b00142.

Van Rossum, G., Drake Jr., F.L., 1995. Python Reference Manual. Centrum voor Wiskunde en Informatica, Amsterdam.

Walser, S.M., Gerstner, D.G., Brenner, B., Höller, C., Liebl, B., Herr, C.E.W., 2014. Assessing the environmental health relevance of cooling towers – a systematic review of legionellosis outbreaks. Int. J. Hyg Environ. Health 217, 145–154. https://doi.org/10.1016/j.ijheh.2013.08.002.

Webster, J.J., Kit, C., 1992. Tokenization as the initial phase in NLP. In: COLING 1992: The 14th International Conference on Computational Linguistics.