Wang *et al. BMC Genomics* (2024) 25:1062
https://doi.org/10.1186/s12864-024-10978-9

BMC Genomics

RESEARCH

Open Access

Interpretation knowledge extraction for genetic testing via question-answer model

Wenjun Wang[1,2,3], Huanxin Chen[1], Hui Wang[4], Lin Fang[4], Huan Wang[5], Yi Ding[6], Yao Lu[4*] and Qingyao Wu[1,3,7]

Abstract

Background Sequencing-based genetic testing is widely used in biomedical research, including pathogenic microorganism detection with metagenomic next-generation sequencing (mNGS). The application of sequencing results to clinical diagnosis and treatment relies on various interpretation knowledge bases. Currently, the existing knowledge bases are primarily built through manual knowledge extraction. This method requires professionals to read extensive literature and extract relevant knowledge from it, which is time-consuming and costly. Furthermore, manual extraction unavoidably introduces subjective biases. In this study, we aimed to automatically extract knowledge for interpreting mNGS results.

Method We propose a novel approach to automatically extract pathogenic microorganism knowledge based on the question-answer (QA) model. First, we construct a MicrobeDB dataset since there is no available pathogenic microorganism QA dataset for training the model. The created dataset contains 3,161 samples from 618 published papers covering 224 pathogenic microorganisms. Then, we fine-tune the selected baseline model based on MicrobeDB. Finally, we utilize ChatGPT to enhance the diversity of training data, and employ data expansion to increase training data volume.

Results Our method achieves an Exact Match (EM) and F1 score of 88.39% and 93.18%, respectively, on the MicrobeDB test set. We also conduct ablation studies on the proposed data augmentation method. In addition, we perform comparative experiments with the ChatPDF tool based on the ChatGPT API to demonstrate the effectiveness of the proposed method.

Conclusions Our method is effective and valuable for extracting pathogenic microorganism knowledge.

Keywords Genetic testing, Interpretation knowledge extraction, Pathogenic microorganism, MicrobeDB, Question-answer

Introduction

Since the advent of next-generation sequencing (NGS) technology, genetic testing has been rapidly developed in the biomedical field, including the clinical detection of pathogenic microorganisms [1, 2]. Raw data of the sequencing results requires bioinformatics analysis and clinical interpretation. At present, mNGS can generate a large amount of genomic data on pathogenic microorganisms in a short period of time and get relatively accurate information about the composition of pathogenic microorganisms through bioinformatics analysis. Interpreting these pathogenic microorganisms is of great

*Correspondence:
Yao Lu
luyaozd@163.com
[1] School of Software Engineering, South China University of Technology, Guangzhou, China
[2] School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang, China
[3] Pazhou Lab, Guangzhou, China
[4] Shenzhen Cladogram Technology Co., Ltd, Shenzhen, China
[5] Industrial Technology Research Center, Guangdong Institute of Scientific & Technical Information, Guangzhou, China
[6] Hunan University of Arts and Science, Changde, China
[7] Peng Cheng Laboratory, Shenzhen, China

significance for clinical diagnosis and treatment [3, 4]. However, in the realm of pathogenic microorganisms, extant databases are primarily biological information databases [5–8]. These resources mainly aid researchers in understanding pathogenic microorganisms' biological characteristics and pathogenic mechanisms. The precise interpretation knowledge base that can assist doctors in clinical diagnosis and treatment is extremely scarce. Therefore, building an accurate pathogenic microorganism interpretation knowledge base is essential to support clinical decision-making. In constructing such a knowledge base, knowledge extraction is crucial. How to accurately and efficiently extract meaningful knowledge that serves the clinic has become an increasingly urgent issue for researchers to consider.

To obtain comprehensive and reliable pathogenic microorganism knowledge that serves clinical decision, we need to collect many research papers on pathogenic microorganisms from diverse sources. At the same time, the data collection process needs to ensure that the collected documents are reliable and closely related to pathogenic microorganisms. Moreover, a team of knowledgeable professionals with a deep understanding of the biological characteristics, pathogenicity, and drug usage of pathogenic microorganisms is required to manually extract the necessary knowledge from research papers and to organize and classify the extracted information in a structured manner.

Currently, knowledge extraction primarily relies on manual methods, as shown in Fig. 1a. Professionals conduct online literature searches to collect relevant papers and extract the required knowledge by manually reading the full text. However, this approach is time-consuming and costly, especially as the number of pathogenic microorganism-related research papers has rapidly increased in recent years. To improve the efficiency of acquiring knowledge, some researchers have adopted automatic online retrieval methods to collect relevant literature and employed simple text processing techniques, such as rule-based keyword highlighting or entity information highlighting based on natural language processing (NLP), to underline potentially important text [9–11]. According to the highlighted text, the required knowledge is manually extracted. Although this method has improved the efficiency of knowledge extraction to some extent, much of the highlighted content is irrelevant. Therefore, it still takes a great deal of time.

To further reduce the time required for manually reading literature and improve the efficiency, in this paper, we propose a novel method for automatically obtaining

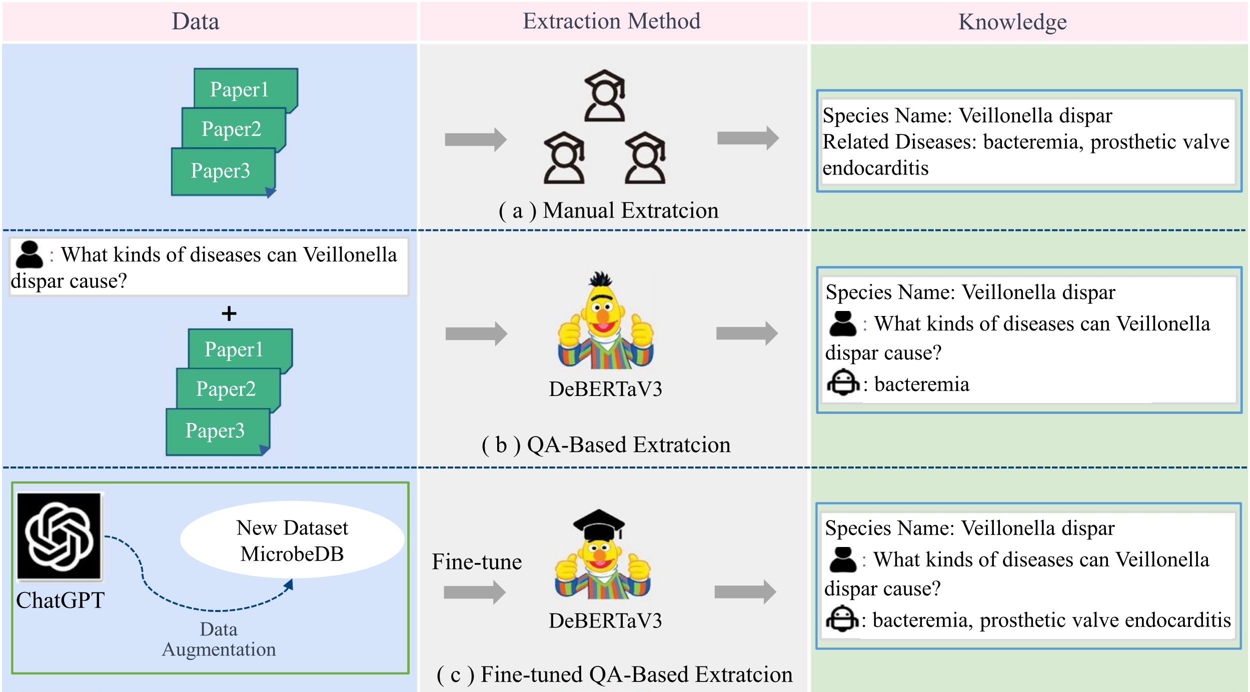

Data
Extraction Method
Knowledge
Paper1
Paper2
Paper3
(a) Manual Extratcion
Species Name: Veillonella dispar
Related Diseases: bacteremia, prosthetic valve endocarditis
: What kinds of diseases can Veillonella dispar cause?
+
Paper1
Paper2
Paper3
DeBERTaV3
(b) QA-Based Extratcion
Species Name: Veillonella dispar
: What kinds of diseases can Veillonella dispar cause?
: bacteremia
ChatGPT
New Dataset MicrobeDB
Data Augmentation
Fine-tune
DeBERTaV3
(c) Fine-tuned QA-Based Extratcion
Species Name: Veillonella dispar
: What kinds of diseases can Veillonella dispar cause?
: bacteremia, prosthetic valve endocarditis

Fig. 1 Comparison of the previous manual method with our proposed method. The proposed approach models knowledge extraction as a QA task, automatically analysing papers and extracting knowledge. We used an existing trained advanced model, such as DeBERTaV3, as our base model. To further improve models' performance, we created the new QA dataset (MicrobeDB) and used ChatGPT to perform data augmentation, which was then used to fine-tune the base model. Best viewed in colour

knowledge about pathogenic microorganisms based on the QA model. Since existing extraction models can answer question accurately, we have modelled knowledge extraction as a QA task, as shown in Fig. 1b. We select DeBERTaV3 [12], one of the latest and most advanced models, and BioBERT [13], specially designed for biomedical text processing, as our knowledge extraction models. However, directly applying these two models to knowledge extraction of pathogenic microorganisms struggles with the problem of task adaptation. To address this, we create a pathogenic microorganism dataset to fine-tune QA models, as shown in Fig. 1c. Our dataset is named MicrobeDB, which contains 3,161 samples drawn from 618 published papers on 224 pathogenic microorganisms. Each question's answer is essential knowledge. Experimental results demonstrate the effectiveness of modelling knowledge extraction as a QA task. To further improve models' performance, we adopt two data augmentation methods. We use ChatGPT and data expansion to increase the diversity and number of training samples. Additionally, we use the ChatPDF tool [14] based on the ChatGPT API to extract pathogenic microorganism knowledge and compare it with our method to show the effectiveness of the proposed approach.

The main contributions of this article are as follows:

- We propose a novel approach to automatically extract pathogenic microorganism knowledge based on the QA model, improving the efficiency of knowledge extraction.
- We create a pathogenic microorganism QA dataset, MicrobeDB, containing 3161 samples by extracting information from 618 published papers related to 224 species of microorganisms.
- To increase the diversity of training data, we employ ChatGPT, the most popular large model, to generate diverse questions, improving the model's performance.
- Extensive experimental results demonstrate the effectiveness of the proposed method.

Related work

Knowledge extraction

Constructing a precise interpretive knowledge base is a complex and time-consuming task that requires the integration of various technologies and methods, such as data collection, database construction, and data mining. At present, public knowledge bases have been established in multiple fields, such as oncology [15] and genetic diseases [16]. Their procedures generally need manually retrieving pertinent papers online from public databases and downloading relevant literature, which is subsequently subjected to manual reading. The extracted knowledge is then stored in a database.

However, manually collecting literature and extracting knowledge is time-consuming and expensive. Particularly, with the rapid growth of related research papers, the cost of human resources will increase exponentially. In recent years, researchers have used online searches to automatically collect relevant literature and adopted some text processing techniques (such as rule-based keyword highlighting or entity information highlighting based on NLP) to underline important text [9–11], improving the efficiency of knowledge extraction. However, this method still requires manual selection and extraction from a large number of candidate texts, thus facing a time-consuming problem. Building a pathogenic microorganism interpretive knowledge base is an even more challenging task. Due to the large number and rapid changes of pathogenic microorganisms and the wide variety of diseases involved, constructing such a knowledge base is more difficult, requiring longer time and higher costs. Recently, Sandra et al. [17] developed an advanced database named Omnicrobe, primarily containing huge descriptions of microbe properties related to food microbe flora. PubMed is the largest source of Omnicrobe and the only source of Omnicrobe for Taxon-Phenotype and Taxon-Use relationships. However, Omnicrobe solely encompasses abstracts sourced from PubMed and does not encompass the remaining sections of papers. Secondly, Omnicrobe's primary audience consists of food processing researchers, agro-industrial technology institutes, agrofood companies, artisans, and food safety agencies, rather than being specifically designed for clinical decision-making. Thirdly, the text mining process for microbial information outlined in Omnicrobe requires three intricate steps: entity recognition, entity normalization, and relation extraction. This process necessitates the integration of information from each step to derive knowledge, and there is a problem of error accumulation [18].

To establish a comprehensive and reliable clinical pathogenic microorganism interpretive knowledge base, a large number of related research papers and high-level professionals are required, and a significant amount of human and financial resources need to be invested. In the biomedical domain, BioASQ organizes the similar task, the challenge on biomedical semantic QA, to facilitate the development of research for extracting information from biomedical text. In BioASQ 2023, prevalent methods [19–23] involve leveraging the strategy of "pre-train and fine-tune" to enhance model performance, suitable for the scenario with limited target training data. Moreover, many approaches [19, 21, 22, 24] have further embraced various data augmentations to enhance the robustness of models. Additionally, some works [20, 25] explore various prompts to enhance the effectiveness of the responses. However, the BioASQ-QA task is for the broad biomedical scientific field and the BioASQ-QA benchmark dataset covers questions as much as possible in medicine, biosciences, and bioinformatics, not specifically for pathogenic microorganism knowledge extraction. In this study, our proposed method is designed for extracting the interpretive knowledge of pathogenic microorganisms. We adopt a "pre-train and fine-tune" training strategy and establish the MicrobeDB dataset to enhance the QA model's adaptability for pathogenic microorganism knowledge extraction. Diverging from the previous data augmentations, we utilize ChatGPT to generate new questions to form additional question-context pairs. Furthermore, the data collection of our method is oriented toward the full texts of retrieved articles.

Question answering

There are two basic types of AI-based QA: extraction QA and generative QA [26]. Extraction QA relies on pre-existing information to extract an answer from a given context. On the other hand, generative QA generates answers relevant to the questions that do not need to come from the original context [27]. Extractive QA is usually suitable for scenarios where both the question and the answer are explicit; generative QA is suitable for scenarios where more ambiguous and open-ended questions are handled. Generative QA is more flexible and detailed in its answers but suffers from some uncertainty. Mistakes and uncertainties might have serious consequences, as the extracted pathogenic microbiological knowledge is for clinical diagnosis and treatment. In contrast, professionals easily check answers predicted using extractive QA to ensure correctness [28]. Therefore, we select the extractive QA methodology. A variety of attention-based interactions between context and query are the early trend [29], including Bidirectional Attention Flow [30], Gated Self-Matching [31], Attention-over-Attention [32] and Fully-Aware Attention [33]. With the advent of BERT [34], QA enters the era of pre-trained models. These pre-trained language models include XLNet [35], RoBERTa [36], T5 [37], BioBERT [13], ALBERT [38], ELECTRA [39], and DeBERTa [12, 40]. Among the models, since DeBERTa is the latest pre-train model and exhibits exceptional performance on a wide range of downstream natural language understanding (NLU) tasks, we select the latest DeBERTaV3 as our baseline model. In addition, since BioBERT is pre-trained on PubMed abstracts and PubMed Central full-text articles and is designed explicitly for biomedical text processing [41, 42], we also add BioBERT as our baseline model.

Materials and methods

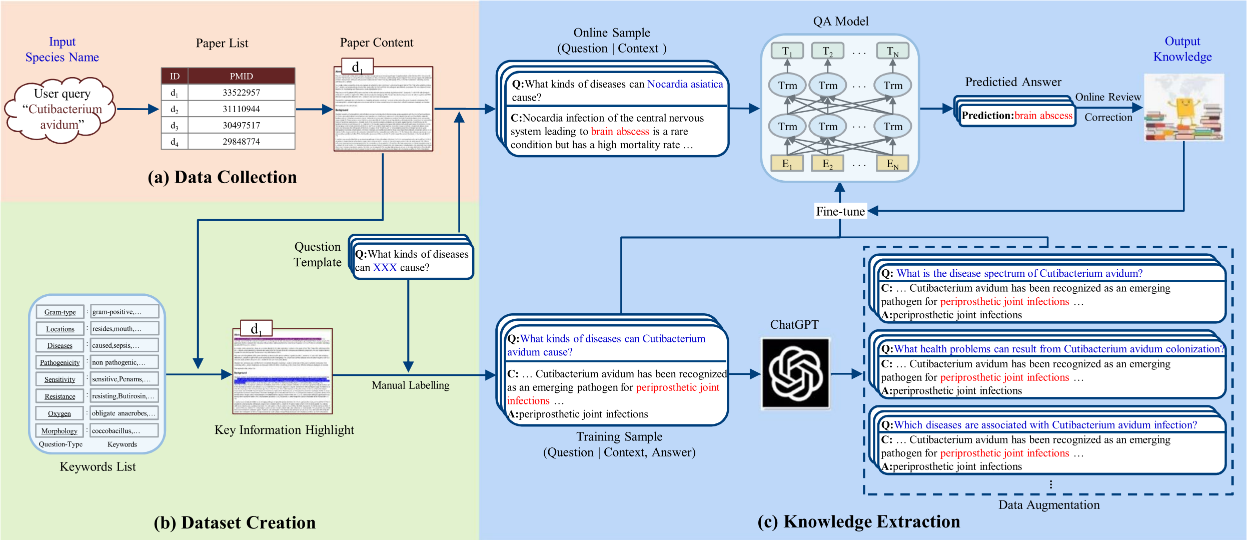
In this section, we first present an overview of our approach for modelling knowledge extraction as a QA task, followed by a description of data collection about pathogenic microorganisms. Subsequently, we introduce dataset creation, producing the MicrobeDB dataset using our annotation tool. Finally, we employ data augmentation through ChatGPT and data expansion to enhance the QA model's performance on pathogenic microorganism knowledge extraction.

Overall framework

Currently, the primary method of extracting pathogenic microorganism knowledge is manual extraction, which is extremely time-consuming and costly. Therefore, we propose a novel approach that combines online retrieval and the QA model to quickly and accurately obtain key knowledge about pathogenic microorganisms. It is a one-stop solution: automatically retrieves relevant literature online, crawls their publicly available contents, and predicts knowledge answers using the QA model. The method consists of data collection, dataset creation, and knowledge extraction, as shown in Fig. 2. For data collection, we utilize the PubMed E-utilities application programming interface for online retrieval and develop a crawler server to obtain publicly available relevant literature content automatically. This can improve efficiency and reduce labour costs. For dataset creation, we first design a pathogenic microorganism knowledge question template. Then we develop an online dataset labelling tool to highlight essential vocabularies and sentences about pathogenic microorganism knowledge, making it easier for labellers to locate crucial information quickly and improving dataset creation efficiency. For knowledge extraction, we select DeBERTaV3 and BioBERT as baseline models and employ two data augmentation methods to improve the models' performance. One is to use the advanced ChatGPT to increase the diversity of training samples. The other is data expansion that reviewed and corrected online samples by professionals can be continually expanded to training samples. The trained model automatically analyses the paper content according to the given knowledge question and predicts the knowledge answer.

Data collection

We develop a data collection tool to enhance the efficacy of gathering data about pathogenic microorganisms. Our tool automatically searches for relevant papers and loads their contents. Specifically, users can first initiate a search on the tool's first page by inputting the desired species name into the designated input box (as exemplified in



ID	PMID
d_1	33522957
d_2	31110944
d_3	30497517
d_4	29848774

Gram-type	gram-positive,…
Locations	resides,mouth,…
Diseases	caused,sepsis,…
Pathogenicity	non pathogenic,…
Sensitivity	sensitive,Penams,…
Resistance	resisting,Butirosin,…
Oxygen	obligate anaerobes,…
Morphology	coccobacillus,…

Q:What kinds of diseases can XXX cause?

Q:What kinds of diseases can Nocardia asiatica cause?
C:Nocardia infection of the central nervous system leading to brain abscess is a rare condition but has a high mortality rate …

Prediction:brain abscess

Q:What kinds of diseases can Cutibacterium avidum cause?
C: … Cutibacterium avidum has been recognized as an emerging pathogen for periprosthetic joint infections …
A:periprosthetic joint infections

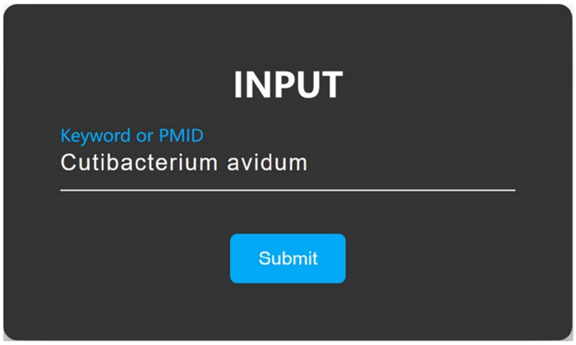
Q: What is the disease spectrum of Cutibacterium avidum?
C: … Cutibacterium avidum has been recognized as an emerging pathogen for periprosthetic joint infections …
A:periprosthetic joint infections

Q:What health problems can result from Cutibacterium avidum colonization?
C: … Cutibacterium avidum has been recognized as an emerging pathogen for periprosthetic joint infections …
A:periprosthetic joint infections

Q:Which diseases are associated with Cutibacterium avidum infection?
C: … Cutibacterium avidum has been recognized as an emerging pathogen for periprosthetic joint infections …
A:periprosthetic joint infections

⋮

Fig. 2 The overall framework of our method. It contains three parts: data collection, dataset creation, and knowledge extraction. **a** We automatically retrieve relevant articles about pathogenic microorganisms and crawl their publicly available contents. **b** We highlight important content based on the keyword list for specific knowledge and manually create a pathogenic microorganism QA dataset. **c** We employ ChatGPT and data expansion to enhance the training samples and subsequently use these samples to fine-tune the QA model to improve the model's performance on knowledge extraction. Best viewed in colour


INPUT

Keyword or PMID
Cutibacterium avidum

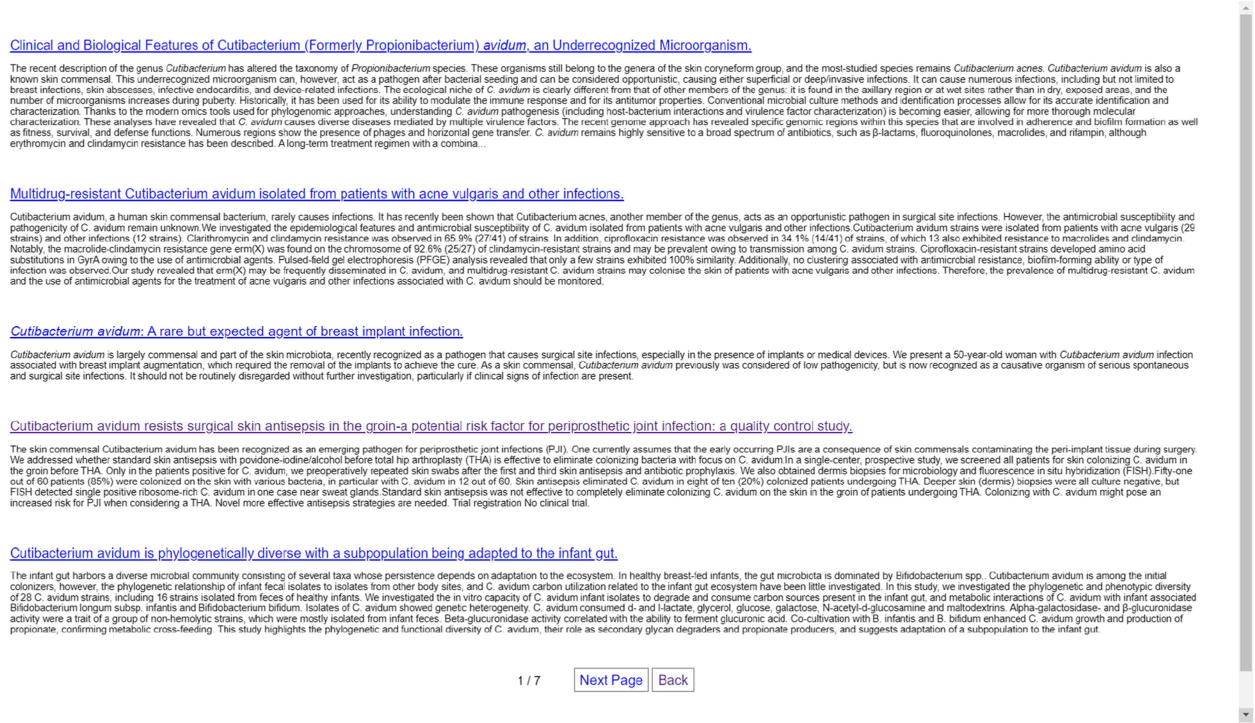
Fig. 3 First page of pathogenic microorganism-related literature search for data collection. This page allows users to enter either a keyword, such as the species Cutibacterium avidum, or a PubMed Identifier (PMID) to search for relevant articles. Best viewed in colour

Fig. 3) and clicking the "Submit" button or pressing the enter key. Then, our tool utilizes the PubMed E-utilities API, provided by the National Center for Biotechnology Information (NCBI), to conduct online searches for papers relevant to the specified species. It retrieves the PubMed Unique Identifier (PMID) lists and presents the corresponding paper titles and their abstracts, as depicted in Fig. 4. When the user selects a publicly available paper or directly enters its PMID, the tool can obtain the article link provided by PubMed and automatically fetch the article's content using our crawler program. Finally, the content is exhibited online, accompanied by a knowledge form area to record the manually extracted knowledge from the paper. The form area is amenable to both manual input of knowledge answers and automatic filling of knowledge answers predicted by the QA model.

Our data collection tool is developed with web technology, utilizing the B/S architecture. The front-end page is rendered through the Vue.js framework, while the MySQL database is used for data storage.

Dataset creation

Although existing advanced QA models have good text comprehension capabilities, they are not tailored for extracting knowledge about pathogenic microorganisms. Directly applying these models to knowledge extraction of pathogenic microorganisms struggles with the problem of task adaptation. As shown in the Results and discussion section (see Baseline models section), they do not yield excellent performance. Therefore, we create a pathogenic microorganism QA dataset for fine-tuning QA models. Due to the specificity of the pathogenic microorganism field, dataset creation necessitates professional annotation, which requires a considerable amount of time to read the full text. We have two dedicated annotators for dataset creation, whose expertise in the relevant domain ensures a nuanced understanding of microbiological information. To improve annotation efficiency, our data collection tool also includes annotation functions. Specifically, we first create a list of keywords



Clinical and Biological Features of Cutibacterium (Formerly Propionibacterium) *avidum*, an Underrecognized Microorganism.

The recent description of the genus *Cutibacterium* has altered the taxonomy of *Propionibacterium* species. These organisms still belong to the genera of the skin coryneform group, and the most-studied species remains *Cutibacterium acnes*. *Cutibacterium avidum* is also a known skin commensal. This underrecognized microorganism can, however, act as a pathogen after bacterial seeding and can be considered opportunistic, causing either superficial or deep/invasive infections. It can cause numerous infections, including but not limited to breast infections, skin abscesses, infective endocarditis, and device-related infections. The ecological niche of *C. avidum* is clearly different from that of other members of the genus: it is found in the axillary region or at wet sites rather than in dry, exposed areas, and the number of microorganisms increases during puberty. Historically, it has been used for its ability to modulate the immune response and for its antitumor properties. Conventional microbial culture methods and identification processes allow for its accurate identification and characterization. Thanks to the modern omics tools used for phylogenomic approaches, understanding *C. avidum* pathogenesis (including host-bacterium interactions and virulence factor characterization) is becoming easier, allowing for more thorough molecular characterization. These analyses have revealed that *C. avidum* causes diverse diseases mediated by multiple virulence factors. The recent genome approach has revealed specific genomic regions within this species that are involved in adherence and biofilm formation as well as fitness, survival, and defense functions. Numerous regions show the presence of phages and horizontal gene transfer. *C. avidum* remains highly sensitive to a broad spectrum of antibiotics, such as β-lactams, fluoroquinolones, macrolides, and rifampin, although erythromycin and clindamycin resistance has been described. A long-term treatment regimen with a combina...

Multidrug-resistant Cutibacterium avidum isolated from patients with acne vulgaris and other infections.

Cutibacterium avidum, a human skin commensal bacterium, rarely causes infections. It has recently been shown that Cutibacterium acnes, another member of the genus, acts as an opportunistic pathogen in surgical site infections. However, the antimicrobial susceptibility and pathogenicity of C. avidum remain unknown.We investigated the epidemiological features and antimicrobial susceptibility of C. avidum isolated from patients with acne vulgaris and other infections.Cutibacterium avidum strains were isolated from patients with acne vulgaris (29 strains) and other infections (12 strains). Clarithromycin and clindamycin resistance was observed in 65.9% (27/41) of strains. In addition, ciprofloxacin resistance was observed in 34.1% (14/41) of strains, of which 13 also exhibited resistance to macrolides and clindamycin. Notably, the macrolide-clindamycin resistance gene erm(X) was found on the chromosome of 92.6% (25/27) of clindamycin-resistant strains and may be prevalent owing to transmission among C. avidum strains. Ciprofloxacin-resistant strains developed amino acid substitutions in GyrA owing to the use of antimicrobial agents. Pulsed-field gel electrophoresis (PFGE) analysis revealed that only a few strains exhibited 100% similarity. Additionally, no clustering associated with antimicrobial resistance, biofilm-forming ability or type of infection was observed.Our study revealed that erm(X) may be frequently disseminated in C. avidum, and multidrug-resistant C. avidum strains may colonise the skin of patients with acne vulgaris and other infections. Therefore, the prevalence of multidrug-resistant C. avidum and the use of antimicrobial agents for the treatment of acne vulgaris and other infections associated with C. avidum should be monitored.

Cutibacterium avidum: A rare but expected agent of breast implant infection.

Cutibacterium avidum is largely commensal and part of the skin microbiota, recently recognized as a pathogen that causes surgical site infections, especially in the presence of implants or medical devices. We present a 50-year-old woman with *Cutibacterium avidum* infection associated with breast implant augmentation, which required the removal of the implants to achieve the cure. As a skin commensal, *Cutibacterium avidum* previously was considered of low pathogenicity, but is now recognized as a causative organism of serious spontaneous and surgical site infections. It should not be routinely disregarded without further investigation, particularly if clinical signs of infection are present.

Cutibacterium avidum resists surgical skin antisepsis in the groin-a potential risk factor for periprosthetic joint infection: a quality control study.

The skin commensal Cutibacterium avidum has been recognized as an emerging pathogen for periprosthetic joint infections (PJI). One currently assumes that the early occurring PJIs are a consequence of skin commensals contaminating the peri-implant tissue during surgery. We addressed whether standard skin antisepsis with povidone-iodine/alcohol before total hip arthroplasty (THA) is effective to eliminate colonizing bacteria with focus on C. avidum.In a single-center, prospective study, we screened all patients for skin colonizing C. avidum in the groin before THA. Only in the patients positive for C. avidum, we preoperatively repeated skin swabs after the first and third skin antisepsis and antibiotic prophylaxis. We also obtained dermis biopsies for microbiology and fluorescence in situ hybridization (FISH).Fifty-one out of 60 patients (85%) were colonized on the skin with various bacteria, in particular with C. avidum in 12 out of 60. Skin antisepsis eliminated C. avidum in eight of ten (20%) colonized patients undergoing THA. Deeper skin (dermis) biopsies were all culture negative, but FISH detected single positive ribosome-rich C. avidum in one case near sweat glands.Standard skin antisepsis was not effective to completely eliminate colonizing C. avidum on the skin in the groin of patients undergoing THA. Colonizing with C. avidum might pose an increased risk for PJI when considering a THA. Novel more effective antisepsis strategies are needed. Trial registration No clinical trial.

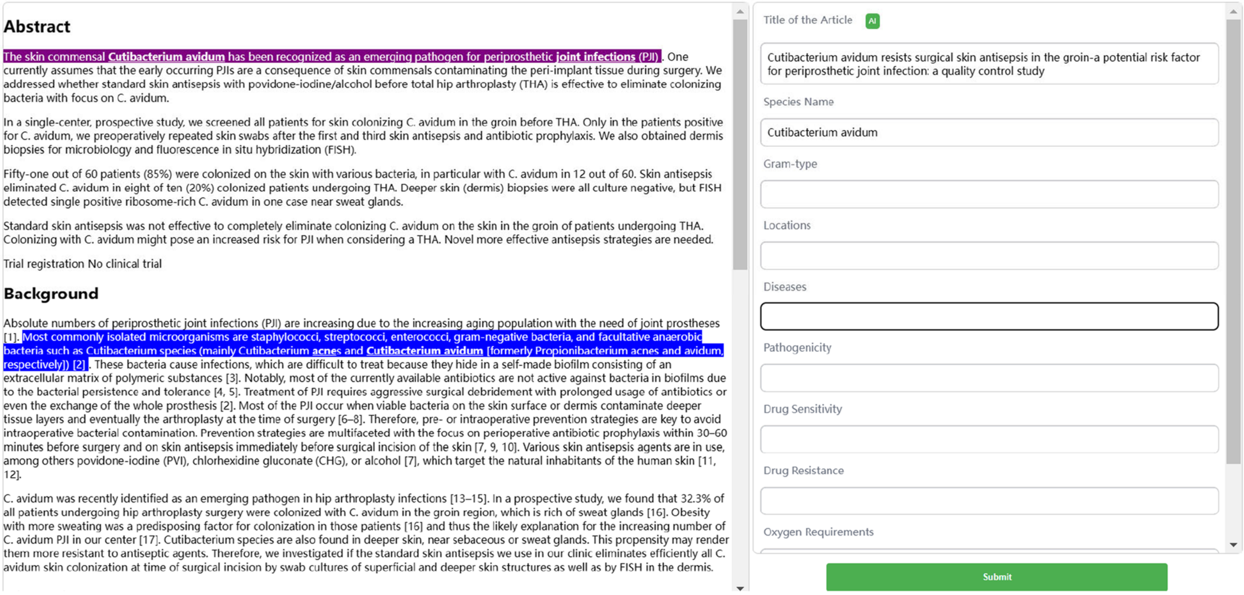
Cutibacterium avidum is phylogenetically diverse with a subpopulation being adapted to the infant gut.

The infant gut harbors a diverse microbial community consisting of several taxa whose persistence depends on adaptation to the ecosystem. In healthy breast-fed infants, the gut microbiota is dominated by Bifidobacterium spp.. Cutibacterium avidum is among the initial colonizers, however, the phylogenetic relationship of infant fecal isolates to isolates from other body sites, and C. avidum carbon utilization related to the infant gut ecosystem have been little investigated. In this study, we investigated the phylogenetic and phenotypic diversity of 28 C. avidum strains, including 16 strains isolated from feces of healthy infants. We investigated the in vitro capacity of C. avidum infant isolates to degrade and consume carbon sources present in the infant gut, and metabolic interactions of C. avidum with infant associated Bifidobacterium longum subsp. infantis and Bifidobacterium bifidum. Isolates of C. avidum showed genetic heterogeneity. C. avidum consumed d- and l-lactate, glycerol, glucose, galactose, N-acetyl-d-glucosamine and maltodextrins. Alpha-galactosidase- and β-glucuronidase activity were a trait of a group of non-hemolytic strains, which were mostly isolated from infant feces. Beta-glucuronidase activity correlated with the ability to ferment glucuronic acid. Co-cultivation with B. infantis and B. bifidum enhanced C. avidum growth and production of propionate, confirming metabolic cross-feeding. This study highlights the phylogenetic and functional diversity of C. avidum, their role as secondary glycan degraders and propionate producers, and suggests adaptation of a subpopulation to the infant gut.

Fig. 4 Results page of pathogenic microorganism-related literature search for data collection. The example displays a list of articles relevant to the species Cutibacterium avidum, including titles and abstracts. Pagination controls are visible at the bottom of the page, allowing navigation through multiple pages of results. Best viewed in colour

corresponding to the desired pathogenic microorganism knowledge, as shown in Supplementary Table S1. The keyword list can be updated, i.e., keywords involved in each type of knowledge can be added, deleted, and modified. Then, when the annotator hopes to annotate specific knowledge, the tool automatically highlights the corresponding keywords and sentences in different colours based on the keyword list for that knowledge, as shown in Fig. 5. With this highlighting method, annotators typically do not need to read the entire text to label answers. If the highlighted content does not provide sufficient information to make an annotation decision, the annotator may proceed to read the relevant context. The labelled results are stored in a database. In research papers, the answers are usually clear. We conduct detailed negotiations on annotation style to maintain consistency. We explicitly specify that annotators should not omit essential modifying words when labeling answers. Moreover, our two annotators, being colleagues working closely together, can communicate seamlessly. They undergo a dual-review process to ensure the dataset's consistency and reliability. When finding a discrepancy annotation, the annotators discuss the reasoning behind their annotations and decide together. If consensus cannot be reached, the sample is marked as a 'fuzzy' sample and is not included in the dataset. These samples are set aside for potential future review and analysis to identify patterns and determine the final annotation. In our practice, no samples have been marked as 'fuzzy' to date, demonstrating the high quality and effectiveness of our annotation process.

When creating the pathogenic microorganism QA dataset, we focus on obtaining clinically relevant knowledge. In order to serve the interpretation of clinical pathogenic microorganisms, the extracted information needs to provide strong support for the diagnosis and treatment of infectious diseases in clinical practice. Through actual communication with clinicians, we emphasize the practicality of question types, ensuring the provision of information with direct guiding significance. Consequently, we identify eight question types, including the pathogen's Gram stain type, locations, related diseases, pathogenicity, drug sensitivity, drug resistance, oxygen requirements, and morphological characteristics. We use the labelled knowledge as the answer, and the question corresponding to the knowledge is taken from the question template, thus constituting the QA pair for each sample. The knowledge question template is shown in Table 1. Specifically, these question types encompass whether the pathogenic microorganism is gram-positive or gram-negative,



Abstract

The skin commensal **Cutibacterium avidum** has been recognized as an emerging pathogen for periprosthetic **joint infections** (PJI). One currently assumes that the early occurring PJIs are a consequence of skin commensals contaminating the peri-implant tissue during surgery. We addressed whether standard skin antisepsis with povidone-iodine/alcohol before total hip arthroplasty (THA) is effective to eliminate colonizing bacteria with focus on C. avidum.

In a single-center, prospective study, we screened all patients for skin colonizing C. avidum in the groin before THA. Only in the patients positive for C. avidum, we preoperatively repeated skin swabs after the first and third skin antisepsis and antibiotic prophylaxis. We also obtained dermis biopsies for microbiology and fluorescence in situ hybridization (FISH).

Fifty-one out of 60 patients (85%) were colonized on the skin with various bacteria, in particular with C. avidum in 12 out of 60. Skin antisepsis eliminated C. avidum in eight of ten (20%) colonized patients undergoing THA. Deeper skin (dermis) biopsies were all culture negative, but FISH detected single positive ribosome-rich C. avidum in one case near sweat glands.

Standard skin antisepsis was not effective to completely eliminate colonizing C. avidum on the skin in the groin of patients undergoing THA. Colonizing with C. avidum might pose an increased risk for PJI when considering a THA. Novel more effective antisepsis strategies are needed.

Trial registration No clinical trial

Background

Absolute numbers of periprosthetic joint infections (PJI) are increasing due to the increasing aging population with the need of joint prostheses [1]. Most commonly isolated microorganisms are staphylococci, streptococci, enterococci, gram-negative bacteria, and facultative anaerobic bacteria such as Cutibacterium species (mainly Cutibacterium **acnes** and **Cutibacterium avidum** [formerly Propionibacterium acnes and avidum, respectively]) [2] . These bacteria cause infections, which are difficult to treat because they hide in a self-made biofilm consisting of an extracellular matrix of polymeric substances [3]. Notably, most of the currently available antibiotics are not active against bacteria in biofilms due to the bacterial persistence and tolerance [4, 5]. Treatment of PJI requires aggressive surgical debridement with prolonged usage of antibiotics or even the exchange of the whole prosthesis [2]. Most of the PJI occur when viable bacteria on the skin surface or dermis contaminate deeper tissue layers and eventually the arthroplasty at the time of surgery [6–8]. Therefore, pre- or intraoperative prevention strategies are key to avoid intraoperative bacterial contamination. Prevention strategies are multifaceted with the focus on perioperative antibiotic prophylaxis within 30–60 minutes before surgery and on skin antisepsis immediately before surgical incision of the skin [7, 9, 10]. Various skin antisepsis agents are in use, among others povidone-iodine (PVI), chlorhexidine gluconate (CHG), or alcohol [7], which target the natural inhabitants of the human skin [11, 12].

C. avidum was recently identified as an emerging pathogen in hip arthroplasty infections [13–15]. In a prospective study, we found that 32.3% of all patients undergoing hip arthroplasty surgery were colonized with C. avidum in the groin region, which is rich of sweat glands [16]. Obesity with more sweating was a predisposing factor for colonization in those patients [16] and thus the likely explanation for the increasing number of C. avidum PJI in our center [17]. Cutibacterium species are also found in deeper skin, near sebaceous or sweat glands. This propensity may render them more resistant to antiseptic agents. Therefore, we investigated if the standard skin antisepsis we use in our clinic eliminates efficiently all C. avidum skin colonization at time of surgical incision by swab cultures of superficial and deeper skin structures as well as by FISH in the dermis.

Title of the Article

Cutibacterium avidum resists surgical skin antisepsis in the groin-a potential risk factor for periprosthetic joint infection: a quality control study

Species Name

Cutibacterium avidum

Gram-type

Locations

Diseases

Pathogenicity

Drug Sensitivity

Drug Resistance

Oxygen Requirements

Fig. 5 An example data annotation page of the species Cutibacterium avidum for dataset creation using our labelling tool. The left side of the figure displays the article content, while the right side shows fields for the article title, species name, and eight question types. When annotators click on the input box under the selected question type, the tool automatically highlights relevant keywords and sentences in different colours based on a keyword list for this question type. In this example, the user selected the input box beneath the Diseases heading, leading the tool to highlight disease-related keywords such as "joint infections" along with the corresponding sentences. Annotators can then label the answer to the question according to the highlighted text. Best viewed in colour

Table 1 Question template per question type

Question type	Question template
Gram	Whether XXX is gram-positive or gram-negative?
Locations	Where does XXX normally exist?
Diseases	What kinds of diseases can XXX cause?
Pathogenicity	What about the pathogenicity of XXX?
Sensitivity	What kinds of drugs are XXX sensitive to?
Resistance	What kinds of drugs are XXX resistant to?
Oxygen	How about XXX's requirement for oxygen?
Morphology	What is the shape of XXX?

'XXX' denotes the species name

how about its requirement for oxygen, where it normally exists, what about its pathogenicity, what diseases it can cause, what drugs it is sensitive to, etc. The answers to these questions can directly assist clinicians in assessing the impact of pathogenic microorganisms in specific environments and guide the formulation of treatment plans. We analyse over 600 published papers involving 224 pathogenic microorganisms. The species included in the dataset are commonly encountered in clinical infections or have been previously reported in infection cases. Ultimately, we produce the MicrobeDB dataset with 3161 samples, with a training-testing split ratio of 7:3, consisting of 2188 and 973 samples, respectively. The number of samples for each question type in the training and test sets is detailed in Table 2. Due to the varying amount of extractable knowledge from each paper, the proportion of samples per question type is different in the dataset. In the process of data collection, we observe that knowledge related to Gram stain type, location, and disease questions is more prevalent in relevant research literature, allowing us to gather a greater number of samples and resulting in a higher proportion of samples for these question types.

Table 2 Detailed distribution of sample numbers for each question type in the MicrobeDB dataset

Dataset	Number of Samples per Question Type								Total
	Gram	Locations	Diseases	Pathogenicity	Sensitivity	Resistance	Oxygen	Morphology	
Training set	327	312	645	160	180	114	209	241	2188
Test set	147	149	310	49	78	41	90	109	973

Knowledge extraction

We select DeBERTaV3 and BioBERT as our knowledge extraction models. The former is one of Microsoft's latest and best-performing pre-trained models. The latter is specifically designed for processing biomedical text and has shown remarkable performance in various NLP tasks within the biomedical field. Though the MicrobeDB dataset contains more than 3100 samples, the number is still relatively small for fine-tuning the two models. Therefore, to further improve their performance, we employ data augmentation through ChatGPT to enhance the diversity of training data and data expansion to increase training data volume.

We use ChatGPT to generate new questions that convey the same meaning but differ in phrasing. We find that there are more different answers to three types of knowledge questions in research papers, including related diseases, locations, and drug sensitivity, as opposed to other questions. Given the complexity of knowledge answers, we enhance the samples of the above three types during model training. Specifically, we employ the generated question, the original context, and the labelled answer to constitute a new sample. For example, for better-extracting knowledge about Moraxella lacunata's related diseases, locations, and drug sensitivity, questions in the template and those generated by ChatGPT (as shown in Table 3) are both utilized for model training. By introducing different question formats, we provide the QA model with diverse samples that help it learn more natural language expressions, improving its accuracy and robustness in pathogenic microorganism knowledge extraction.

In addition, we use data expansion to increase the number of training samples. Specifically, we first use the MicrobeDB training set and its augmentation through ChatGPT to fine-tune the QA model. The fine-tuned QA model is then employed to predict the answers for the specific knowledge questions in online papers, and these predictions are highlighted in conjunction with their corresponding sentences to facilitate professionals checking the response's correctness quickly. When the extracted knowledge is wrong, they correct the results online. The reviewed and corrected answers and the corresponding questions are also used to further fine-tune the QA model to help it learn more features. This approach allows us to continually increase new training samples, improving the model's performance.

Results and discussion

In this section, we conduct extensive experiments to demonstrate the effectiveness of the proposed method. Firstly, we present the experiments of DeBERTaV3 and BioBERT on the MicrobeDB dataset. Subsequently, we employ data augmentation methods to improve their performance. Finally, we perform comparative experiments on knowledge extraction between the proposed method and ChatPDF.

Table 3 Enhancing the knowledge question diversity on Moraxella lacunata's related diseases, locations, and drug sensitivity through ChatGPT

Question template for related diseases	What kinds of diseases can Moraxella lacunata cause?
Enhancing question diversity about related diseases	What is the disease spectrum of Moraxella lacunata?
	What health problems can result from Moraxella lacunata colonization?
	Which diseases are associated with Moraxella lacunata infection?
	What are the diseases that can be caused by Moraxella lacunata?
	What types of illnesses can Moraxella lacunata contribute to?
Question template for locations	Where does Moraxella lacunata normally exist?
Enhancing question diversity about locations	What are the typical habitats of Moraxella lacunata?
	In what environments can Moraxella lacunata be found?
	What are the common sites where Moraxella lacunata is known to inhabit?
	Where is Moraxella lacunata commonly present?
	In what locations can Moraxella lacunata usually be found?
Question template for drug sensitivity	What kinds of drugs are Moraxella lacunata sensitive to?
Enhancing question diversity about drug sensitivity	Which drugs are effective against Moraxella lacunata?
	What medications can be used to treat Moraxella lacunata infections?
	What drugs have been shown to be active against Moraxella lacunata?
	What are the drugs that Moraxella lacunata is vulnerable to?
	Which antibiotics are recommended for treating Moraxella lacunata infections?

Baseline models

In this section, we conduct experiments using the MicrobeDB dataset to evaluate two baseline models for pathogenic microorganism knowledge extraction: DeBERTaV3 and BioBERT. Both are pre-trained language models based on the Transformer [43] architecture and have been extensively fine-tuned to adapt well to the QA task. We select the trained weight files provided by Hugging Face for DeBERTaV3 and BioBERT: deberta-v3-base-squad2 [44] and biobert_v1.1_pubmed_squad_v2 [45], respectively, and discover that they do not yield excellent performance for pathogenic microorganism knowledge extraction. Therefore, to better adapt DeBERTaV3 and BioBERT to this task, we fine-tune the models using the MicrobeDB training set. We use Exact Match (EM) and F1 score as performance evaluation metrics, as with most extractive QA tasks [46]. The experimental results are listed in Table 4. This table shows that using the MicrobeDB dataset greatly improves models' performance, indicating the importance of creating a pathogenic microorganism QA dataset.

Furthermore, Table 4 demonstrates that the fine-tuned DeBERTaV3 performs comparably to the fine-tuned BioBERT. Their EM values are the same, and the F1 score of DeBERTaV3 is 1.02% lower than BioBERT. This may be due to BioBERT being pre-trained on biomedical domain corpora (PubMed abstracts and PMC full-text articles). When the number of samples for fine-tuning is insufficient, BioBERT can better adapt to the specific language and knowledge of the biomedical field, resulting in better performance. Since professionals pay more attention to the exact matching of predicted and labelled answers, we use both models as baseline models for subsequent experiments.

Data augmentation

In this study, we utilize data augmentation to enhance the performance of QA models. Given the already high performance of these fine-tuned baseline models, achieving further improvements is challenging and meaningful. We employ data augmentation through ChatGPT to increase the diversity of training data, aiming to improve models' generalization ability and robustness. We find that there are more different answers to three types of knowledge questions in research papers, including related diseases, locations, and drug sensitivity, as opposed to other questions. Therefore, we select to enhance the three types of knowledge questions. For each of the above types, ChatGPT is utilized to generate five additional questions that are semantically similar to those in the question template but have different phrasing. This augmentation approach is plug-and-play, requiring minimal adjustment to integrate with existing models. From Table 5, we observe that the performance of both baseline models has improved. Specifically, the DeBERTaV3 model has demonstrated a 1.44% increase in EM and a 1.09% increase in F1 score, resulting in 88.18% and 93.14%, respectively. Similarly, the BioBERT model has shown a 0.93% increase in EM and a 0.05% increase in F1 score, reaching 87.67% and 93.12%, respectively. These are non-negligible improvements in this specialized field. We augment the training samples with diverse question variations to enhance the model's ability to understand different expressions. For example, in disease-related questions, the augmented samples included different expressions such as 'contribute to,' 'result from,' and 'associate with.' These variations help the models learn to extract answers from different contexts, regardless of how the information is phrased in those contexts. Consequently, after applying ChatGPT augmentation, prediction errors in disease-related questions decreased obviously–from 62 to 52 errors in the DeBERTaV3 model, and from 60 to 55 errors in the BioBERT model. This increased diversity enables the models to handle different sentence structures more effectively, thereby improving their accuracy in contextual knowledge extraction. The experimental results indicate that using ChatGPT to augment the disease and location questions with diversity is effective, except for drug sensitivity questions. We discover that many drug names of drug sensitivity questions appearing in

Table 4 Performance comparison of the fine-tuned DeBERTaV3 model and the fine-tuned BioBERT model on the MicrobeDB test dataset

QA model	Number of prediction errors per question type								EM	F1 score
	Gram	Locations	Diseases	Pathogenicity	Sensitivity	Resistance	Oxygen	Morphology		
Pretrained DeBERTaV3	126	71	191	39	37	15	64	90	34.94	47.68
Fine-tuned DeBERTaV3	4	21	62	8	15	4	2	13	**86.74**	**92.05**
Pretrained BioBERT	114	61	158	45	36	15	78	98	37.82	50.06
Fine-tuned BioBERT	2	20	60	10	20	8	1	8	**86.74**	**93.07**

The best performance is highlighted in bold

Table 5 Results of ablation experiments on two data augmentation methods

QA model	Number of prediction errors per question type								EM	F1 score
	Gram	**Locations**	**Diseases**	**Pathogenicity**	**Sensitivity**	**Resistance**	**Oxygen**	**Morphology**		
Baseline(DeBERTaV3)	4	21	62	8	15	4	2	13	86.74	92.05
+ChatGPT augmentation	3	20	52	7	16	3	2	12	88.18	93.14
+ChatGPT augmentation+training set expansion	2	19	46	10	15	5	3	13	**88.39**	**93.18**
Baseline(BioBERT)	2	20	60	10	20	8	1	8	86.74	93.07
+ChatGPT augmentation	2	17	55	7	20	6	2	11	87.67	93.12
+ChatGPT augmentation+training set expansion	2	15	56	8	20	5	2	6	**88.28**	**93.32**

The best performance is highlighted in bold while the second-best performance is underlined pathogenic microbiology-related papers are outside the token table of QA models, which affects the understanding of these out-of-vocabulary (OOV) drug words, phrases and contexts, making it difficult to achieve good answers. Therefore, we plan to add new drug-related tokens to the token table of QA models to improve the models' understanding of drug-related contexts and overall performance. Our work demonstrates the effectiveness of ChatGPT-driven augmentation in this task. In the future, we will explore more large model prompts, experiment with different large models, and investigate other large model-based augmentation methods to achieve better results.

In addition, we employ data expansion to improve the model's performance. After generating the model's predictions for knowledge answers online, professionals can review the predictions. When the extracted answers are wrong, they correct the results online. The reviewed and corrected answers are then integrated with the contexts and related questions, producing new samples. These online samples are subsequently used to fine-tune QA models. This method allows for the continuous addition of additional training samples. We expand the training set with 466 new samples obtained from 89 papers through this method, including 78 samples for Gram stain type, 58 samples for locations, 157 samples for related diseases, 19 samples for pathogenicity, 42 samples for drug sensitivity, 41 samples for drug resistance, 22 samples for oxygen requirements, and 49 samples for morphological characteristics. Experimental results are shown in Table 5. DeBERTaV3 reached 88.39% and 93.18% in EM and F1 score, respectively, and BioBERT arrived at 88.28% and 93.32% in EM and F1 score, respectively. Both these pre-augmentation and post-augmentation results demonstrate the effectiveness of modelling knowledge extraction as a QA task.

Compare with ChatPDF

ChatPDF is a tool using the ChatGPT API for quickly extracting the needed information from any PDF file [47]. Upon receiving a user query, ChatPDF presents the relevant paragraphs and the question to the text-generation model and returns the generated answer to the user. We also utilize ChatPDF to extract the desired knowledge from research papers on pathogenic microorganisms. We randomly choose 25 papers covering 17 species and pose 143 questions that could be answered using the information within these papers. The predicted answer to the question is valuable if it is correct. We compare ChatPDF with our method. Since the DeBERTaV3 model, fine-tuned using the MicrobeDB training set and the above two data augmentations, has the best EM value in the MicrobeDB test set, we employ the fine-tuned DeBERTaV3 as the QA model used in our approach. The detailed comparative record is presented in Supplementary Table S2. We find it necessary to occasionally modify the questions asked to obtain the correct answers when using ChatPDF. The experimental results show that our method has 1 question where one of the answers is incorrect, but ChatPDF has 13 questions where it replies that the provided article does not contain the answer information, i.e., none of the answers to these knowledge questions are extracted.

Furthermore, compared to ChatPDF, we discover that the proposed method is also superior in the following two ways: First, the answers extracted by our method are directly from the original text and are more accurate and concise. It is easier to facilitate the traceability of answers and verify the accuracy of answers, making it suitable for clinical decision-making services. Moreover, the answers from ChatPDF are sometimes extended with irrelevant information. It is difficult to judge the accuracy of answers generated by ChatPDF. Second, our approach allows for the batch processing of all knowledge questions for a paper, automatically loading and displaying the article's content and quickly predicting their answers. The answers can be submitted with one click after the review is completed and saved to the backend database, facilitating efficient data management. However, ChatPDF does not provide batch processing to the questions and needs manual questioning, manual copying of answers, and manual data management, which is troublesome. These findings suggest that our method is more suitable than ChatPDF for extracting pathogenic microorganism knowledge.

Explore the effect of sample variation on model performance

To understand the effect of variations in sample size for different question types on model performance, we employ a specific strategy: we keep the number of samples constant for seven question types during each model training and halve the sample quantity for one specific question type. Thus, a total of eight distinct models are trained, and a comprehensive evaluation is conducted, as presented in Table 6. The fine-tuned DeBERTaV3 model on the full training set with the best EM value is used as the comparative model. From Table 6, it can be observed that for half of question types, such as locations, diseases, etc., a reduction in the training sample quantity for the specific question type led to a decline in the model's predictive performance on that question. For the other half of question types, the testing results are similar to the outcomes of the comparative model. This shows that more training samples may have a positive impact on model performance.

Table 6 The effect of variations in the number of samples for different question types on model performance

Question type with halve in sample size	Number of prediction errors per question type								EM	F1 score
	Gram	Locations	Diseases	Pathogenicity	Sensitivity	Resistance	Oxygen	Morphology		
-	2	19	46	10	15	5	3	13	88.39	93.18
Gram	**4**	22	51	9	17	7	1	11	87.46	92.31
Locations	4	**22**	54	8	15	3	2	10	87.87	92.58
Diseases	4	23	**60**	8	15	6	2	12	86.64	92.10
Pathogenicity	3	19	59	**9**	16	7	2	12	86.95	92.43
Sensitivity	4	22	57	7	**17**	6	3	11	86.95	92.28
Resistance	3	21	56	10	15	**5**	2	12	87.26	92.26
Oxygen	3	20	56	8	18	5	**2**	11	87.36	92.67
Morphology	3	22	53	8	20	5	2	**13**	87.05	91.85

Furthermore, we conduct an evaluation on a novel set of samples to assess the generalizability of our approach. Our evaluation encompasses a new dataset featuring previously unseen question types related to clinical infections, such as how about the virulence of the pathogenic microorganism, whether it has catalase, how about its motility, whether it forms spores, etc. The dataset contains 247 samples. Notably, our approach achieves 83.81% accuracy (EM) and an 89.32% F1 score.

Conclusion

In this study, we propose a novel method for automatically extracting knowledge about pathogenic microorganisms based on the QA model. Our method automatically retrieves related articles, crawls their publicly available contents, and predicts the answers to the targeted knowledge. To make the QA model better suitable for pathogenic microorganism knowledge extraction, we create the MicrobeDB dataset to fine-tune it. Moreover, we utilize ChatGPT to enhance the diversity of training data and employ data expansion to increase training data volume. Extensive experiments demonstrate the effectiveness of the proposed method. However, currently there are two limitations to our approach. First, the MicrobeDB dataset only includes eight question types related to pathogenic microorganisms. This may lead to suboptimal performance when the model encounters the new knowledge extraction task for new question types. Our future plan is to incorporate more question types related to interpretation of clinical pathogenic microorganisms and improve the model effectiveness to better serve clinical decision-making. Second, our database contains redundant knowledge extracted from different papers. In the future, we will consider how to remove redundant knowledge to eventually build a comprehensive and reliable knowledge base.

Abbreviations

mNGS	Metagenomic next-generation sequencing
PMID	PubMed unique identifier
NCBI	National center for biotechnology information
PMC	PubMed central
NLP	Natural language processing
EM	Exact match
NLU	Natural language understanding
OOV	Out-of-vocabulary

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10978-9.

Supplementary Material 1.

Acknowledgements

Not applicable.

Authors' contributions

WW: methodology, software, writing - original draft, writing - review and editing. HC: software, writing - review and editing. HW: data curation, writing - review and editing. LF: investigation, data curation. HW: methodology, analysis. YD: analysis, funding acquisition. YL: project administration, funding acquisition. QW: conceptualization, writing – review and editing. All authors have read and approved the manuscript.

Funding

This work was supported by National Natural Science Foundation of China (NSFC) 62272172 and 62266012, Guangzhou Basic Research Program-Basic and Applied Basic Research Project 2023A04J1051, Guangdong Institute of Scientific and Technical Information Strategic Studies Project 2023B003, and the High-Level Innovative Talent Project of Guizhou Province (Grant no. QKHPTRC-GCC2023027).

Data availability

The MicrobeDB dataset, knowledge extraction code, and trained model weights, generated during the current study, are available in the MicrobeQAExtractor repository, https://doi.org/10.5281/zenodo.13981795 or https://github.com/WenjunWang-SCUT/MicrobeQAExtractor.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 10 July 2023 Accepted: 29 October 2024
Published online: 09 November 2024

References

1. Leung CM, Li D, Xin Y, Law WC, Zhang Y, Ting HF, et al. MegaPath: sensitive and rapid pathogen detection using metagenomic NGS data. BMC Genomics. 2020;21(6):1–9.
2. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G, Committee PP, et al. Validation of metagenomic next-generation sequencing tests for universal pathogen detection. Arch Pathol Lab Med. 2017;141(6):776–86.
3. Simner PJ, Miller S, Carroll KC. Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases. Clin Infect Dis. 2018;66(5):778–88.
4. Dulanto Chiang A, Dekker JP. From the pipeline to the bedside: advances and challenges in clinical metagenomics. J Infect Dis. 2020;221(Supplement_3):S331–S340.
5. Hu R, Yao R, Li L, Xu Y, Lei B, Tang G, et al. A database of animal metagenomes. Sci Data. 2022;9(1):312.
6. Wu L, Sun Q, Desmeth P, Sugawara H, Xu Z, McCluskey K, et al. World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Res. 2017;45(D1):D611–8.
7. Federhen S. The NCBI taxonomy database. Nucleic Acids Res. 2012;40(D1):D136–43.
8. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinforma. 2019;20(4):1125–36.
9. Xu Q, Liu Y, Hu J, Duan X, Song N, Zhou J, et al. OncoPubMiner: a platform for mining oncology publications. Brief Bioinforma. 2022 09;23(5).
10. Lee S, Kim D, Lee K, Choi J, Kim S, Jeon M, et al. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. PLoS ONE. 2016;11(10):e0164680.
11. Allot A, Peng Y, Wei CH, Lee K, Phan L, Lu Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Res. 2018;46(W1):W530–6.
12. He P, Gao J, Chen W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In: Proceedings of the 11th International Conference on Learning Representations. Kigali: ICLR; 2023.
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
14. und Moritz Lage GbR ML. Chat with any PDF. 2023. https://www.chatpdf.com/. Accessed 27 Apr 2023.
15. Holt ME, Mittendorf KF, LeNoue-Newton M, Jain NM, Anderson I, Lovly CM, et al. My cancer genome: coevolution of precision oncology and a molecular oncology knowledgebase. JCO Clin Cancer Inform. 2021;5:995–1004.
16. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh AOMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(D1):D789–98.
17. Dérozier S, Bossy R, Deléger L, Ba M, Chaix E, Harlé O, et al. Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. PLoS ONE. 2023;18(1):e0272473.
18. Han J, Wang H. Transformer based network for open information extraction. Eng Appl Artif Intell. 2021;102:104262.
19. Aksenova A, Asamov T, Ivanov P, Boytcheva S. Improving Biomedical Question Answering with Sentencebased Ranking at BioASQ-11b. In: CEUR Workshop Proceedings, vol. 3497. Thessaloniki: CEUR-WS; 2023. p. 27–36.
20. Nentidis A, Katsimpras G, Krithara A, Lima López S, Farré-Maduell E, Gasco L, et al. Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and question answering. In: Proceedings of the 14th International Conference of the Cross-Language Evaluation Forum for European Languages. Thessaloniki: Springer; 2023. p. 227–50.
21. KKim H, Hwang H, Lee C, Seo M, Yoon W, Kang J. Exploring Approaches to Answer Biomedical Questions: From Pre-processing to GPT-4 Notebook for the BioASQ Lab at CLEF 2023. In: CEUR Workshop Proceedings, vol. 3497. Thessaloniki: CEUR-WS; 2023. p. 132–44.
22. Panou D, Reczko M. Semi-Supervised Training for Biomedical Question Answering. In: CEUR Workshop Proceedings, vol. 3497. Thessaloniki: CEUR-WS; 2023. p. 152–8.
23. Galat D, Rizoiu MA. Enhancing Biomedical Text Summarization and Question-Answering: On the Utility of Domain-Specific Pre-Training University of Technology Sydney participation in BioASQ Task 11b Phase B. In: CEUR Workshop Proceedings, vol. 3497. Thessaloniki: CEUR-WS; 2023. p. 102–13.
24. Mitamura T. Biomedical Question Answering with Transformer Ensembles. In: CEUR Workshop Proceedings, vol. 3497. Thessaloniki: CEUR-WS; 2023. p. 159–67.
25. Hsueh CY, Zhang Y, Lu YW, Han JC, Meesawad W, Tsai RTH. NCU-IISR: Prompt Engineering on GPT-4 to Stove Biological Problems in BioASQ 11b Phase B. In: CEUR Workshop Proceedings, vol. 3497. Thessaloniki: CEUR-WS; 2023. p. 114–21.
26. Luo M, Hashimoto K, Yavuz S, Liu Z, Baral C, Zhou Y. Choose Your QA Model Wisely: A Systematic Study of Generative and Extractive Readers for Question Answering. In: 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge, Spa-NLP 2022. Dublin: ACL; 2022. p. 7–22.
27. Xu X, Tohti T, Hamdulla A. A Survey of Machine Reading Comprehension Methods. In: 2022 International Conference on Asian Language Processing (IALP). Singapore: IEEE; 2022. p. 312–17.
28. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL; 2016. p. 2383–92.
29. Zhang Z, Yang J, Zhao H. Retrospective reader for machine reading comprehension. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, vol. 16. virtual: AAAI; 2021. p. 14506–14.
30. Seo M, Kembhavi A, Farhadi A, Hajishirzi H. Bi-directional Attention Flow for Machine Comprehension. In: Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR; 2017.
31. Wang W, Yang N, Wei F, Chang B, Zhou M. Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL; 2017. p. 189–98.
32. Cui Y, Chen Z, Wei S, Wang S, Liu T, Hu G. Attention-over-Attention Neural Networks for Reading Comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL; 2017. p. 593–602.
33. Huang HY, Zhu C, Shen Y, Chen W. FusionNet: Fusing via Fully-aware Attention with Application to Machine Comprehension. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR; 2018.
34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL; 2019. p. 4171–86.
35. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. Vancouver: Neural information processing systems foundation; 2019. p. 5753–63.
36. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692; 2019.
37. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.

38. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR; 2020.
39. Clark K, Luong MT, Le QV, Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR; 2020.
40. He P, Liu X, Gao J, Chen W. Deberta: Decoding-enhanced bert with disentangled attention. In: Proceedings of the 9th International Conference on Learning Representations. Virtual: ICLR; 2021.
41. Chen Y. A transfer learning model with multi-source domains for biomedical event trigger extraction. BMC Genomics. 2021;22:1–18.
42. Dholakia D, Kalra A, Misir BR, Kanga U, Mukerji M. HLA-SPREAD: a natural language processing based resource for curating HLA association from PubMed abstracts. BMC Genomics. 2022;23:1–14.
43. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach: Neural information processing systems foundation; 2017. p. 5999–6009.
44. Lee S, Möller T, Malte P. deepset/deberta-v3-base-squad2. 2023. https://huggingface.co/deepset/deberta-v3-base-squad2. Accessed 16 Apr 2023.
45. Kirill T. ktrapeznikov/biobert_v1.1_pubmed_squad_v2. 2021. https://huggingface.co/ktrapeznikov/biobert_v1.1_pubmed_squad_v2. Accessed 16 Apr 2023.
46. Lewis P, Oguz B, Rinott R, Riedel S, Schwenk H. MLQA: Evaluating Cross-lingual Extractive Question Answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual: ACL; 2020. p. 7315–30.
47. Li J, Tanabe H, Ota K, Gu W, Hasegawa S. Automatic Summarization for Academic Articles using Deep Learning and Reinforcement Learning with Viewpoints. In: Proceedings of the 36th International Florida Artificial Intelligence Research Society Conference, vol. 36. Clearwater Beach: Florida Online Journals; 2023.

Publisher's Note