Rapid response to contamination of drinking water distribution networks

Paraskevopoulos, S.

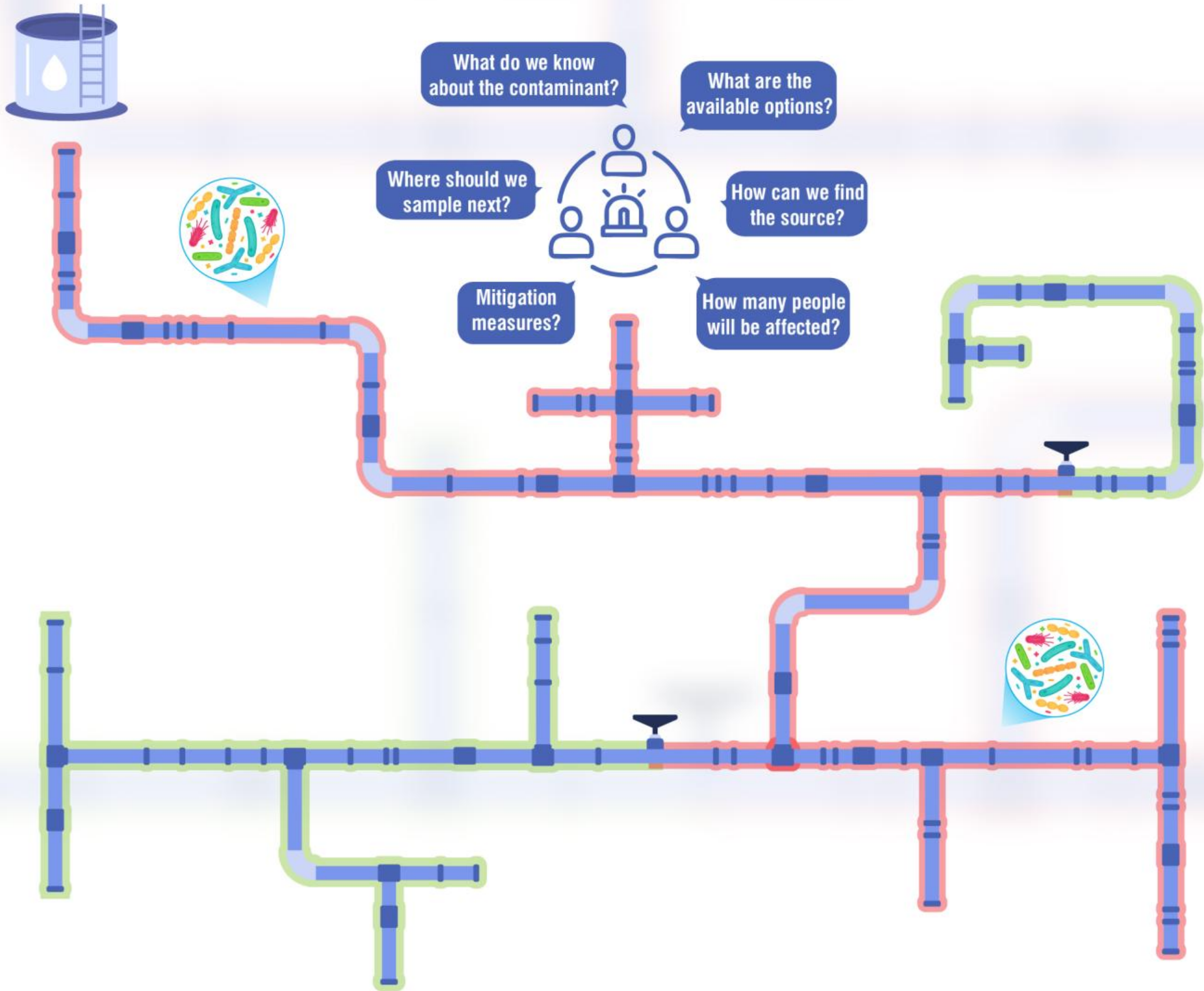
DOI
10.4233/uuid:266d81af-e781-4570-99f7-674521bc7629

Publication date
2025

Document Version
Final published version

Citation (APA)
Paraskevopoulos, S. (2025). *Rapid response to contamination of drinking water distribution networks*. [Dissertation (TU Delft), Delft University of Technology]. Delft University of Technology. https://doi.org/10.4233/uuid:266d81af-e781-4570-99f7-674521bc7629

Rapid response to contamination of drinking water distribution networks



Sotirios Paraskevopoulos

“Rapid response to contamination of drinking water distribution networks”

Dissertation

for the purpose of obtaining the degree of doctor at Delft University of Technology
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates to be defended publicly on
Friday 31 October 2025 at 10:00 o'clock

by

Sotirios PARASKEVOPOULOS

Master of Science in Earth Sciences, Utrecht University, the Netherlands
born in Larisa, Greece

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof. dr. G. J. Medema	Delft University of Technology, promotor
Prof. dr. Z. Kapelan	Delft University of Technology, promotor
Dr.ir. P. W. M. H. Smeets	KWR Water Research Institute, NL, External advisor

Independent members:

Prof. dr. ir. P.H.A.J.M. van Gelder	Delft University of Technology
Prof.dr. A.M. de Roda Husman	Utrecht University/RIVM, NL
Prof.dr. D.A. Savic	University of Exeter, UK
Prof.dr. M. Prévost	Polytechnique Montreal, Canada

Reserve member:

Prof.dr.ir. D. van Halem	Delft University of Technology

DOI 10.4233/uuid:266d81af-e781-4570-99f7-674521bc7629

ISBN 978-94-6384-849-7

“Rapid response to contamination of drinking water distribution networks”

Proefschrift

ter verkrijging van de graad van doctor aan de Technische Universiteit Delft
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties, in het openbaar te verdedigen op
Friday 31 October 2025 at 10:00 o'clock

by

Sotirios PARASKEVOPOULOS

Master of Science in Earth Sciences, Utrecht University, the Netherlands
geboren te Larisa, Griekenland

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie bestaat uit:

Rector magnificus,	voorzitter
Prof. dr. G. J. Medema	Delft University of Technology, promotor
Prof. dr. Z. Kapelan	Delft University of Technology, promotor
Dr.ir. P. W. M. H. Smeets	KWR Water Research Institute, NL, Externe adviseur

Onafhankelijke leden:

Prof. dr. ir. P.H.A.J.M. van Gelder	Delft University of Technology
Prof.dr. A.M. de Roda Husman	Utrecht University/RIVM, NL
Prof.dr. D.A. Savic	University of Exeter, UK
Prof.dr. M. Prévost	Polytechnique Montreal, Canada

Reservelid:

Prof.dr.ir. D. van Halem	Delft University of Technology

DOI 10.4233/uuid:266d81af-e781-4570-99f7-674521bc7629

ISBN 978-94-6384-849-7

Summary

The provision of safe drinking water is essential in every society since it determines people's health and well-being. Drinking Water Distribution Networks (DWDN) are vital for this purpose but are susceptible to pathogen contamination and outbreaks due to cascading events after infrastructure failures, main repairs, human errors, or malicious attacks. When a contamination event occurs in the DWDN, the preservation of health of the public should be the top priority in every emergency response mechanism. Exposure to contaminated water can cause significant health risks by introducing pathogens such as enterovirus, *Campylobacter*, and *Cryptosporidium.* For this reason, DWDNs are nowadays considered critical infrastructures, recognized by USA's Presidential Policy Directive 21 and the European Union's Directive (EU) 2022/2557.

During contamination events in the DWDN, water utilities need to act quickly, make informed decisions, assess the threat, and effectively mitigate the event. The central objective of this thesis was to generate knowledge to help address the growing challenge of waterborne pathogen contamination in DWDNs and develop applications that can enhance decision-making and immediate actions in such emergencies. Tools and methodologies were developed and evaluated focusing on two main pillars. The first pillar involves understanding the event based on historical knowledge. Innovative approaches were developed and assessed for Artificial Intelligence-based information extraction and question-answering using scientific publications, enabling rapid access to up-to-date pathogen characteristics, historical information on contamination events, and control actions. The second pillar focuses on predicting and managing the specific contamination event in real-time. Advanced modeling tools were created to simulate contamination events in DWDNs, providing realistic representations of hydraulics and water quality dynamics, predicted health impacts, and support for real-time decision-making during emergencies.

Chapter 2 describes the development of an Artificial Intelligence (AI)-based model that extracts specific pathogen information from the scientific literature. By leveraging Natural Language Processing (NLP) and Deep Learning (DL) techniques, the study evaluated the feasibility and performance of an Information Extraction model to extract both qualitative and quantitative information from scientific publications about the waterborne pathogen *Legionella*. For the development of the model, a combination of supervised and rule-based techniques was adopted. The evaluation metrics showed a satisfactory performance for extraction of both qualitative and quantitative information with an overall F-score of 85% and 95% for the supervised and rule-based technique respectively. The model was also compared with a human extraction, returning similar results and indicating that the extracted information is of high quality. The results showed that the model can be used to rapidly extract critical information from text documents and be a useful tool for water utilities, enabling faster and more informed decision-making during the early stages of contamination.

Chapter 3 systematically assesses the performance of various open-source Large Language Models (LLM), including Llama 2, Mistral, and Gemma (and their variations) in a question-answering task related to pathogen contamination events of drinking water. The evaluation metrics included Precision, Recall, F1 score, Automated Accuracy, and Empty Score. The model with the highest performance on a set of 23 questions using 188 scientific publications was then manually evaluated by a human (Human Evaluation). The results showed that all models performed reasonably well with an average F1 score ranging from 81% to 87%. After considering all the evaluation metrics, the Llama 2 model was the most reliable model with an average Automated Accuracy of 86%. However, the hallucination effect of Llama 2 was evident. The Gemma model had a lower Automated Accuracy score but was less prone to hallucination. The Human Evaluation showed that the Llama 2 model delivered correct answers when the questions were clear and straightforward. However, when the question required further interpretation, the model often struggled. Overall, the study demonstrated that the use of LLMs in automated information extraction tasks show great potential for time-critical applications, such as processing large volumes of (historical) data in real-time thereby making it feasible to make historical information available in near real-time in case of emergencies.

Building on the response to a pathogen contamination event in the DWDN, **Chapter 4** presents the BeWaRE benchmark testbed, a comprehensive model. This testbed went beyond the state-of-the art and integrated all current relevant knowledge on pathogen transport and fate, bulk and wall chlorine decay, fast and slow chlorine reactions with TOC, TOC degradation, stochastic water demands, hydraulic uncertainty, and individual consumption patterns to calculate pathogen exposure and infection risk following the steps of Quantitative Microbial Risk Assessment (QMRA). A large wastewater contamination in different locations in a chlorinated and non-chlorinated network was simulated using three pathogens: *Campylobacter*, enterovirus, and *Cryptosporidium*. The results of this study showed that in non-chlorinated DWDNs, the modeled wastewater contamination event led to 11-46% infection risk in the total population, depending on the contamination location, but irrespective of the selected pathogen (due to the high pathogen concentration). On the other hand, in chlorinated DWDNs, the same scenarios resulted in lower infection risk for the pathogens that are susceptible to chlorine; 0.78-2.1 % for *Campylobacter* and 7.8-26.6 % for enterovirus. Moreover, the enterovirus infection risk was higher, despite the concentrations in the contamination source being lower, due to the lower susceptibility to chlorine than *Campylobacter*. While chlorination aids mitigation, large contaminations can still lead to infections due to chlorine resistance (for Cryptosporidium) and chlorine depletion at the contamination point. Finally, the varying levels of pathogen susceptibility to chlorine, the contamination location and duration, influenced the infection risk, while the response window to reduce the health impact was short; in these scenarios 5-10 hours post-contamination. The study provided a novel approach to assessing health risks, offering critical insights for water utilities to optimize their response during emergencies.

Chapter 5 further explores the added value of using modeling tools to support decision-making during emergencies in the DWDN. This was demonstrated through PathoINVEST, an analytical tool that utilizes the BeWaRE benchmark methodology, which was presented in the previous Chapter, to support water utilities in modeling contamination events in the DWDN. A case study was conducted with the aim of comparing a traditional approach (representing the status quo of current practices of water utilities) with a model-based approach (use of real-time modeling tools) during an emergency response to a contamination event in the DWDN. The model-based approach was shown to be more efficient than the traditional approach in identifying the source of contamination (1.3 versus 3.7 hours), requiring fewer samples (4 versus 11) and resulting in lower infection risk by the time the source was identified (12% versus 20%) in this case study. Moreover, the model-based approach was more effective in finding the best valves to close in the network (as mitigation measures) since it resulted in a 3%-point infection risk reduction. However, some actions taken in the traditional approach, such as the rapid closure of valves (cutting the network in half and thus limiting further spreading) before the contamination source was identified, were critical in mitigating the contamination. Another key finding was the importance of having an up-to-date overview of valve settings in the DWDN schematization to provide reliable results on source identification since any discrepancies between the actual network and the model can lead to inaccurate infection risk estimates when using modeling tools to support decision-making. Overall, this case study showed that integrating modeling tools in the current practices of water utilities provides a robust framework for improving water contamination management and decision-making processes, thus safeguarding public health during emergencies.

A concluding viewpoint is offered in **Chapter 6**, which considers whether the initial research questions from Chapter 1 were successfully answered. The implications of this research for water utilities are examined, providing information on how the proposed methodologies can be (and have been) used in real-world scenarios, facilitating a faster decision-making and contributing to effective mitigation of emergencies. Finally, the perspectives and future research are discussed, emphasizing the role of AI and the advancements in modeling tools. AI has shown significant potential in enhancing situational awareness and rapid information extraction during emergencies. Water utilities should explore the integration of AI into their standard operating procedures to further enhance emergency responses and routine management. Regarding the use of modeling tools during emergencies, future research should address key gaps, such as the complex dynamics when wastewater interacts with chlorine, the competition between chlorine-reducing agents, and the validity of hydraulic modeling assumptions such as perfect mixing. Accounting for cumulative health risks (multiple pathogens) and refining dose-response models to differentiate between infection and illness probabilities can provide insights for effectively managing risks to vulnerable populations. Moreover, the incorporation of metrics like Disability-Adjusted Life Years (DALYs) into modeling efforts could enable better communication of health impacts and evaluation of mitigation strategies. Finally, Digital Twins and real-time microbial sensors are identified as transformative technologies that can provide real-time insights into network dynamics. These advancements can shift water utility management from reactive approaches to proactive, data-driven strategies, significantly enhancing public health protection, operational efficiency, and resilience.

Samenvatting

Het leveren van veilig drinkwater is essentieel in elke samenleving, omdat het de gezondheid en het welzijn van mensen bepaalt. Drinkwaterdistributienetwerken (DWDN) zijn van vitaal belang voor dit doel, maar zijn vatbaar voor besmetting met ziekteverwekkers en uitbraken als gevolg van opeenvolgende gebeurtenissen na infrastructuurstoringen, hoofdreparaties, menselijke fouten of kwaadaardige aanvallen. Wanneer er een besmettingsgebeurtenis plaatsvindt in het DWDN, moet het behoud van de volksgezondheid de hoogste prioriteit hebben in elk noodresponsmechanisme. Blootstelling aan besmet water kan aanzienlijke gezondheidsrisico's veroorzaken door de introductie van ziekteverwekkers zoals enterovirus, Campylobacter en Cryptosporidium. Om deze reden worden DWDN's tegenwoordig beschouwd als kritieke infrastructuren, erkend door de Presidential Policy Directive 21 van de VS en de Richtlijn (EU) 2022/2557 van de Europese Unie.

Tijdens besmettingsgebeurtenissen in het DWDN moeten waterbedrijven snel handelen, weloverwogen beslissingen nemen, de dreiging beoordelen en de gebeurtenis effectief beperken. De centrale doelstelling van de studie van dit proefschrift was om kennis te genereren om de groeiende uitdaging van besmetting met watergedragen pathogenen in DWDN's aan te pakken en toepassingen te ontwikkelen die besluitvorming en onmiddellijke acties in dergelijke noodsituaties kunnen verbeteren. Er werden hulpmiddelen en methodologieën ontwikkeld en geëvalueerd met de focus op twee hoofdpijlers. De eerste pijler omvat het begrijpen van de gebeurtenis op basis van historische kennis. Er werden innovatieve benaderingen ontwikkeld en beoordeeld voor op kunstmatige intelligentie gebaseerde informatie-extractie en vraagbeantwoording met behulp van wetenschappelijke publicaties, waardoor snelle toegang tot actuele kenmerken van pathogenen, historische informatie over besmettingsgebeurtenissen en controleacties mogelijk werd. De tweede pijler richt zich op het voorspellen en beheren van de specifieke besmettingsgebeurtenis in realtime. Er werden geavanceerde modelleringshulpmiddelen gecreëerd om besmettingsgebeurtenissen in DWDN's te simuleren, die realistische representaties van hydrauliek en waterkwaliteitsdynamiek, voorspelde gezondheidseffecten en ondersteuning voor realtime besluitvorming tijdens noodsituaties bieden.

Hoofdstuk 2 beschrijft de ontwikkeling van een op kunstmatige intelligentie (AI) gebaseerd model dat specifieke pathogeeninformatie uit de wetenschappelijke literatuur haalt. Door gebruik te maken van Natural Language Processing (NLP) en Deep Learning (DL)-technieken, evalueerde de studie de haalbaarheid en prestaties van een Information Extraction-model om zowel kwalitatieve als kwantitatieve informatie te halen uit wetenschappelijke publicaties over de watergedragen pathogeen Legionella. Voor de ontwikkeling van het model werd een combinatie van supervised en rule-based technieken gebruikt. De evaluatiemetrieken lieten een bevredigende prestatie zien voor extractie van zowel kwalitatieve als kwantitatieve informatie met een algehele F-score van respectievelijk 85% en 95% voor de supervised en rule-based techniek. Het model werd ook vergeleken met een menselijke extractie, wat vergelijkbare resultaten opleverde en aangaf dat de geëxtraheerde informatie van hoge kwaliteit is. De resultaten lieten zien dat het model kan worden gebruikt om snel kritieke informatie uit tekstdocumenten te halen en een nuttig hulpmiddel kan zijn voor waterbedrijven, waardoor snellere en beter geïnformeerde besluitvorming mogelijk is in de vroege stadia van besmetting.

Hoofdstuk 3 beoordeelt systematisch de prestaties van verschillende open-source Large Language Models (LLM), waaronder Llama 2, Mistral en Gemma (en hun variaties) in een vraag-en-antwoordtaak met betrekking tot pathogene besmettingsgebeurtenissen van drinkwater. De evaluatiemetrieken omvatten Precisie, Herinnering, F1-score, Geautomatiseerde nauwkeurigheid en Lege score. Het model met de hoogste prestatie op een set van 23 vragen met behulp van 188 wetenschappelijke publicaties werd vervolgens handmatig geëvalueerd door een mens (Menselijke evaluatie). De resultaten toonden aan dat alle modellen redelijk goed presteerden met een gemiddelde F1-score variërend van 81% tot 87%. Na overweging van alle evaluatiemetrieken was het Llama 2-model het meest betrouwbare model met een gemiddelde Geautomatiseerde nauwkeurigheid van 86%. Het hallucinatie-effect van Llama 2 was echter duidelijk. Het Gemma-model had een lagere Geautomatiseerde nauwkeurigheidsscore, maar was minder vatbaar voor hallucinaties. De menselijke evaluatie toonde aan dat het Llama 2-model correcte antwoorden gaf wanneer de vragen duidelijk en eenvoudig waren. Echter, wanneer de vraag

verdere interpretatie vereiste, had het model vaak moeite. Over het geheel genomen toonde de studie aan dat het gebruik van LLM's in geautomatiseerde informatie-extractietaken een groot potentieel heeft voor tijdkritische toepassingen, zoals het verwerken van grote volumes (historische) data in real-time, waardoor het haalbaar wordt om historische informatie in near rea-time beschikbaar te maken in geval van noodgevallen.

Voortbouwend op de respons op een besmettingsgebeurtenis met een pathogeen in het DWDN, presenteert **Hoofdstuk 4** de BeWaRE benchmark testbed, een uitgebreid model. Deze testbed ging verder dan de state-of-the-art en integreerde alle huidige relevante kennis over pathogeentransport en -bestemming, bulk- en wandchloorverval, snelle en langzame chloorreacties met TOC, TOC-afbraak, stochastische waterbehoeften, hydraulische onzekerheid en individuele consumptiepatronen om blootstelling aan pathogenen en infectierisico te berekenen volgens de stappen van Quantitative Microbial Risk Assessment (QMRA). Een grote afvalwaterverontreiniging op verschillende locaties in een gechloreerd en niet-gechloreerd netwerk werd gesimuleerd met behulp van drie pathogenen: Campylobacter, enterovirus en Cryptosporidium. De resultaten van deze studie toonden aan dat in niet-gechloreerde DWDN's de gemodelleerde afvalwaterverontreiniging leidde tot een infectierisico van 11-46% in de totale populatie, afhankelijk van de besmettingslocatie, maar ongeacht de geselecteerde pathogeen (vanwege de hoge concentratie pathogeen). Aan de andere kant resulteerden dezelfde scenario's in gechloreerde DWDN's in een lager infectierisico voor de pathogenen die gevoelig zijn voor chloor; 0,78-2,1% voor Campylobacter en 7,8-26,6% voor enterovirus. Bovendien was het enterovirusinfectierisico hoger, ondanks dat de concentraties in de besmettingsbron lager waren, vanwege de lagere vatbaarheid voor chloor dan Campylobacter. Hoewel chlorering helpt bij het verminderen, kunnen grote besmettingen nog steeds leiden tot infecties vanwege chloorresistentie (voor Cryptosporidium) en chlooruitputting op het besmettingspunt. Ten slotte beïnvloedden de verschillende niveaus van vatbaarheid van pathogenen voor chloor, de locatie en duur van de besmetting het infectierisico, terwijl het responsvenster om de gezondheidsimpact te verminderen kort was; in deze scenario's 5-10 uur na besmetting. De studie bood een nieuwe benadering voor het beoordelen van gezondheidsrisico's en bood cruciale inzichten voor waterbedrijven om hun respons tijdens noodsituaties te optimaliseren.

Hoofdstuk 5 onderzoekt verder de toegevoegde waarde van het gebruik van modelleringshulpmiddelen ter ondersteuning van besluitvorming tijdens noodsituaties in het DWDN. Dit werd aangetoond via PathoINVEST, een analytisch hulpmiddel dat gebruikmaakt van de BeWaRE-benchmarkmethodologie, die in het vorige hoofdstuk werd gepresenteerd, om waterbedrijven te ondersteunen bij het modelleren van besmettingsgebeurtenissen in het DWDN. Er werd een case study uitgevoerd met als doel een traditionele aanpak (die de status quo van de huidige praktijken van waterbedrijven vertegenwoordigt) te vergelijken met een op modellen gebaseerde aanpak (gebruik van realtime modelleringshulpmiddelen) tijdens een noodrespons op een besmettingsgebeurtenis in het DWDN. De op modellen gebaseerde aanpak bleek efficiënter te zijn dan de traditionele aanpak bij het identificeren van de bron van besmetting (1,3 versus 3,7 uur), waarbij minder monsters nodig waren (4 versus 11) en het infectierisico daalde op het moment dat de bron werd geïdentificeerd (12% versus 20%) in deze case study. Bovendien was de op modellen gebaseerde aanpak effectiever in het vinden van de beste kleppen om te sluiten in het netwerk (als mitigerende maatregelen), omdat het resulteerde in een vermindering van het infectierisico met 3%. Sommige acties die in de traditionele aanpak werden ondernomen, zoals het snel sluiten van kleppen (het netwerk doormidden snijden en zo verdere verspreiding beperken) voordat de bron van de besmetting werd geïdentificeerd, waren echter cruciaal bij het beperken van de besmetting. Een andere belangrijke bevinding was het belang van een actueel overzicht van klepinstellingen in de DWDN-schematisering om betrouwbare resultaten te bieden bij het identificeren van bronnen, omdat eventuele discrepanties tussen het werkelijke netwerk en het model kunnen leiden tot onnauwkeurige schattingen van het infectierisico bij het gebruik van modelleringshulpmiddelen ter ondersteuning van de besluitvorming. Over het geheel genomen toonde deze casestudy aan dat het integreren van modelleringshulpmiddelen in de huidige praktijken van waterbedrijven een robuust raamwerk biedt voor het verbeteren van het beheer van waterverontreiniging en besluitvormingsprocessen, waardoor de volksgezondheid tijdens noodsituaties wordt beschermd.

Een afsluitend standpunt wordt geboden in **Hoofdstuk 6**, waarin wordt bekeken of de initiële onderzoeksvragen uit Hoofdstuk 1 succesvol zijn beantwoord. De implicaties van dit onderzoek voor waterbedrijven worden onderzocht, waarbij informatie wordt verstrekt over hoe de voorgestelde methodologieën kunnen worden (en zijn) gebruikt in real-world scenario's, wat een snellere besluitvorming mogelijk maakt en bijdraagt aan effectieve mitigatie van noodsituaties. Tot slot worden de perspectieven en toekomstig onderzoek besproken, waarbij de nadruk ligt op de rol van AI en de vooruitgang in modelleringshulpmiddelen. AI heeft een aanzienlijk potentieel getoond bij het verbeteren van situationeel bewustzijn en snelle informatie-extractie tijdens noodsituaties. Waterbedrijven moeten de integratie van AI in hun standaardwerkprocedures onderzoeken om noodresponsen en routinematig beheer verder te verbeteren. Met betrekking tot het gebruik van modelleringshulpmiddelen tijdens noodsituaties moet toekomstig onderzoek belangrijke hiaten aanpakken, zoals de complexe dynamiek wanneer afvalwater interageert met chloor, de concurrentie tussen chloorreducerende middelen en de geldigheid van hydraulische modelleringsaannames zoals perfecte menging. Rekening houden met cumulatieve gezondheidsrisico's (meerdere pathogenen) en het verfijnen van dosis-responsmodellen om onderscheid te maken tussen infectie- en ziektewaarschijnlijkheden kan inzicht bieden in het effectief beheren van risico's voor kwetsbare bevolkingsgroepen. Bovendien kan de opname van statistieken zoals Disability-Adjusted Life Years (DALY's) in modelleringsinspanningen betere communicatie van gezondheidseffecten en evaluatie van mitigatiestrategieën mogelijk maken. Tot slot worden Digital Twins en realtime microbiële sensoren geïdentificeerd als transformatieve technologieën die realtime inzicht kunnen bieden in netwerkdynamiek. Deze ontwikkelingen kunnen het beheer van waterbedrijven verschuiven van reactieve benaderingen naar proactieve, datagestuurde strategieën, wat de bescherming van de volksgezondheid, operationele efficiëntie en veerkracht aanzienlijk verbetert.

Abbreviations

AI= Artificial Intelligence

BattLeDIM =Battle of the Leakage Detection and Isolation Methods

BERT= Bidirectional Encoder Representations from Transformers

BeWaRE= Benchmark for Water network and Risk Evaluation

CFU= Colony Forming Units

CRA= Chlorine-reducing agents

CRF= Conditional Random Fields

CSI= Contamination Source Identification

DL= Deep Learning

DW= Drinking Water

DWDN= Drinking Water Distribution Network

EHR= Electronic Health Records

EU= European Union

FRA= Fast reducing agents

GA= Genetic Algorithm

GPT= Generative Pre-trained Transformer

IE= Information Extraction

IK= Information Keywords

IWS= Intermittent Water Supply

LLMs= Large Language Models

MeSH= Medical Subject Headings

ML= Machine Learning

NCWS= Non-Community Water Systems

NER= Named Entity Recognition

NLP= Natural Language Processing

NOM= Natural Organic Matter

PathoCERT= Pathogen Contamination Emergency Response Technologies

PathoINVEST= Pathogen contamination INVESTigation decision support system

PFU= Plaque forming units

POC= Proof of Concept

PRISMA= Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PSO= Particle Swarm Optimization

PWS= Public Water Systems

QA= Question-answering

QMRA= Quantitative Microbial Risk Assessment

RCT= Randomized Control Trials

Regex= Regular expressions

RNN= Recurrent Neural Networks

RPTB= Response Protocol Toolbox

RT-PCR= Reverse Transcription Polymerase Chain Reaction

SRA= Slow reducing agents

STREaM= Stochastic Residential water End-use Model

SVM= Support Vector Machine

TM= Text Mining

TOC= Total Organic Carbon

UMLS= Unified Medical Language System

Table of Contents

Chapter 1

1.1 Introduction

1.1.1 Drinking water networks and contamination events

The provision of safe drinking water (DW) is essential in every society since it determines people's health and well-being. Drinking Water Distribution Networks (DWDN) are vital for this purpose but are susceptible to pathogen contamination and outbreaks due to cascading events after natural disasters, infrastructure failures, main repairs, human errors, or malicious attacks (Hrudey and Hrudey, 2004; Fewtrell et al., 2011; Cann et al., 2013; Blokker et al., 2018).

A study by Risebro et al. (2007) indicated that DWDN deficiencies (mainly backflow and cross-connections) can significantly contribute to waterborne outbreaks, indicating their severe impact. In 19 of the 61 outbreaks analyzed, these deficiencies were typically solitary but catastrophic due to the absence of barriers between the incident and the consumers. They concluded that emphasis should be made to enhanced monitoring, maintenance, and effective communication between relevant stakeholders to reduce the risk of such incidents.

When a contamination event occurs in the DWDN, the preservation of health of the public should be the top priority in every emergency response mechanism (Erickson et al., 2019). The exposure to contaminated water bodies can cause significant health risks by introducing pathogens such as enterovirus, *Campylobacter*, and *Cryptosporidium* into the water supply.

Notable incidents highlight the vulnerability of those systems to contamination. For example, the Milwaukee cryptosporidiosis outbreak (Mac Kenzie et al., 1995) affected more than 400,000 people. In Nokia, Finland a wastewater contamination resulted in 8,453 cases (Laine et al., 2011). The cryptosporidiosis infection in the town of Skellefteå, Sweden had 27,000 cases (Bjelkmar et al., 2017). Additionally, an outbreak in Antwerp, Belgium by surface water infiltration during firefighting (Braeye et al., 2015) resulted in 222 cases, while wastewater infiltration events in Italy (Giammanco et al., 2018) and in Denmark (Kuhn et al., 2017) resulted in 25 and 63 cases respectively. Finally, recently (August 2024) there have been reports of suspected tampering with the water supply at a military base in Germany[1]. For this reason, they are nowadays considered critical infrastructures, recognized by USA's Presidential Policy Directive 21 and the European Union's Directive (EU) 2022/2557.

There are different types of DWDN settings with different types of water sources around the world. For example, in the United States, the system is categorized into Public Water Systems (PWS), including Community Water Systems (CWS) and Non-Community Water Systems (NCWS) with more than 91% of these systems using groundwater as their primary water source (Craun and Calderon, 2001). While CWS serves the same population all year long, NCWS serves temporary populations such as schools and factories. On the other hand, in Europe, excluding small-scale water supplies typically found in rural areas (WHO, 2011), a more centralized and uniform approach is taken, with a balanced water source usage. For example, in the Netherlands the water source usage is 60% groundwater versus 40% surface water (Geudens and Grootveld, 2017). Regardless of the DWDN setting, or the water source, numerous waterborne disease outbreaks associated with DWDN deficiencies have happened over the last 50 years both in Europe (Table 1.1) and in US (Table 1.2) with multiple illness cases and hospitalizations.

---

[1] Nostlinger and Lau (2024, August 14). German authorities suspect water supply sabotage on military base. https://www.politico.eu/article/water-supply-sabotage-military-bases-germany-nato-cologne-geilenkirchen/

Table 1.1 Overview of microbiological contamination events in the DWDNs of Europe. Table modified from Tangena, 2018.

Location	Year	Cause	Cases	Hospitalizations	Deaths
Rotterdam, NL	1981	human error	609	0	0
Leidsche Rijn, NL	2001	human error	37	0	0
Haarlemmermeer, NL	2007	technical failure	0	0	0
Freuchie, SCH	1995	human error	765	5	0
Greater Belfast,NlE	2001	human error	306	41	0
Santa Maria de Palautordera,ES	2002	human error	756	14	0
Bergen, NO	2004	technical failure	2,500	0	0
NW Wales,GB	2005	technical failure	231	0	0
Oslo, NO	2007	technical failure	0	0	0
Nokia, Fl	2007	human error	8,453	204	0
Galway,IE	2007	technical failure	304	40	0
Northamptonshire	2008	technical failure	422	0	0
Adliswil, CH	2008	human error	180	0	0
Ostersund, SE	2010	technical failure	27000	51	0
Koge, DK	2010	technical failure	409	0	0
Hemiksem/Schelle, BE	2010	human error	603	6	0
Copenhagen, DK	2011	naturally	0	0	0
Ronse/Kluisbergen, BE	2013	unknown	0	0	0
Lancashire, GB	2015	technical failure	2	0	0
Praag, CZ	2015	technical failure	150	0	0

*Table 1.2 Waterborne disease outbreaks in PWSs caused by DWDN deficiencies and water system sources for the period 1971–1998 in the USA. *Outbreaks per 10000 water systems. ** Milwaukee, Wis., cryptosporidiosis outbreak (403,000 cases; 440 hospitalizations; 50 deaths) excluded from this analysis. Table modified from Craun and Calderon, 2001.*

	Groundwater		**Surface water**					
System type	**Outbreaks**	**Rate***	**Outbreaks**	**Rate**	**Cases**	**Cases/Outbreak**	**Hospitalizations**	**Deaths**
CWS**	36	9	33	62.6	17220	193.5	488	13
NCWS	17	1.6	1	5.4	3838	159.9	10	0
PWS	53	3.4	34	26.2	21058	186.4	498	13

1.1.2 Water utilities and emergency response

A meta-analysis of a systematic literature review of 188 scientific publications related to pathogen contamination events and outbreaks of drinking water (DW) worldwide revealed that 86% of the time, the contamination was detected only after a critical window of opportunity, typically through unusual increases in hospitalizations[2]. Figure 1.1 shows that only 7% of the outbreaks were identified as the event was unfolding from water utilities due to customer complaints.

---

[2] The systematic literature review was performed for the purposes of the publication in Chapter 3.

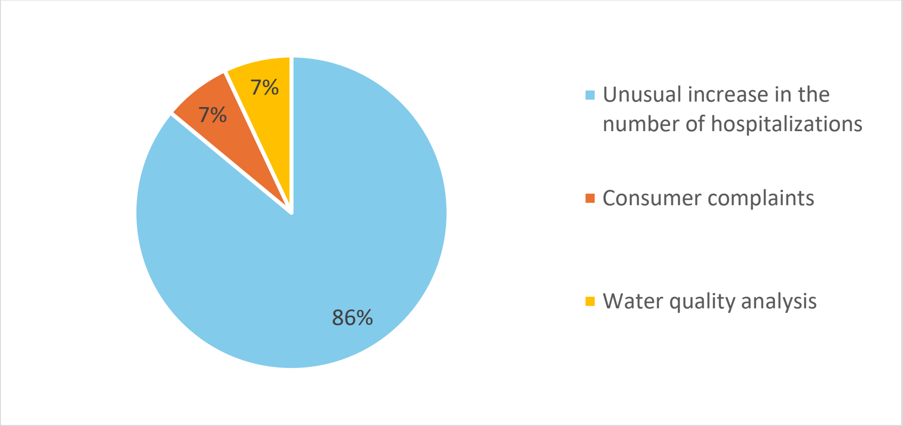


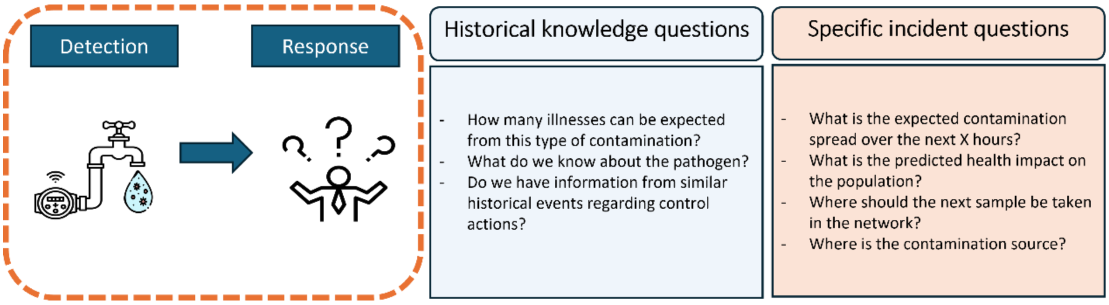
Figure 1.1 Results of the meta-analysis of the literature review regarding how pathogen contamination events and outbreaks have been detected.

In the Netherlands, several contaminations per year are detected after routine water quality monitoring in DWDNs. Although a boil water advice is following, these events are not recorded as outbreaks. This indicates that the actual number of contamination events is much higher than the reported outbreaks. As shown in Table 1.1, several contaminations have happened without recorded illness cases. Similarly, Craun and Calderon (2011) reported that only 10-30 % of US waterborne disease outbreaks are reported, highlighting the variability in outbreak detection and reporting.

These statistics (Figure 1.1) highlight the imperative need to develop tools for real-time water quality monitoring of DWDNs. Moreover, the low 7% detection through complaints (during the event) highlights the importance of not missing the rare opportunities to respond rapidly. In these occasions, water utilities need to act quickly, make informed decisions, assess the threat, and effectively mitigate the event.

The type of questions typically formed during an emergency response focus on specific, actionable information that can directly inform and ideally improve decision-making and immediate actions. These questions can be grouped into two main pillars that reflect the critical needs of water utilities during such events (Figure 1.2):

1. First Pillar: Understanding the event based on historical knowledge
2. Second Pillar: Predicting and managing the specific incident in real-time



- How many illnesses can be expected from this type of contamination?
- What do we know about the pathogen?
- Do we have information from similar historical events regarding control actions?

- What is the expected contamination spread over the next X hours?
- What is the predicted health impact on the population?
- Where should the next sample be taken in the network?
- Where is the contamination source?

Figure 1.2. The different types of questions a water utility has during a pathogen contamination event in their network.

1.1.2.1 First pillar

Status quo

When water utilities experience pathogen contamination events in their network, the available information on what they are facing at that time is often limited. They typically rely on established operational procedures and existing knowledge, which may not include the up-to-date information and global knowledge on pathogens and past -similar- contamination events. Since contamination events are (fortunately) rare, experience in contamination event response at utility level is usually limited.

Although every situation is unique and a tailored response is necessary, the knowledge that is available from past outbreaks and contamination events can be highly beneficial to deduce a knowledge-based response. This information needs to come fast to respond adequately and protect public health. The type of information needed includes pathogen characteristics (e.g., incubation time, symptoms, fate), potential impacts (e.g., expected hospitalizations and deaths from similar past events), and mitigation measures (e.g., chlorination, boil water advisories). Traditionally, acquiring any type of information involves time-consuming literature reviews that are not only slow but also require specific expertise.

The lack of access to up-to-date information and the need for prompt responses underscore the necessity of developing specialized tools that can be used from the very beginning of a pathogen contamination event in the DWDN. Artificial intelligence (AI), providing rapid data processing and information extraction capabilities, could provide a robust solution for rapidly accessing information from a constantly increasing volume of text and scientific publications on pathogens and contamination events in the DWDN.

1.1.2.2 Second pillar

Status quo

When it comes to the tailored, specific response to a pathogen contamination event in a DWDN, the affected water utility wants to assess health impacts and respond quickly and effectively. Traditional decision-making during these crises has been predominantly guided by past experience, standard operating procedures, emergency response protocols, and expert judgment based on (limited) available information on the contamination event and knowledge of network characteristics. Experts using network modeling tools may be consulted, but with the current state of the network modeling of contamination events, their input often comes after a critical window of opportunity (sometimes after 24 hours have passed since the event detection). They frequently have to rely on outdated network characteristics and slower, generic (not dedicated to pathogen contamination events) models, without being able to rapidly model the propagation and infection risk of contamination. Since the response cannot wait, actions will be taken by the utility response team, but without the proper intelligence and models, this can lead to inefficient and ineffective actions that unintendedly increase the health impact of a contamination event.

A realistic representation of contamination type and site, demand-driven hydraulics, understanding of contaminant transport, and the effect of a residual disinfectant and valve manipulation as mitigation measures are crucial for a comprehensive risk assessment and rapid, efficient response. Moreover, the integration of new (real-time) modeling tools in their current (traditional) practices can enhance responses and reduce the negative impacts of a contamination event.

1.1.3 The need for change

Despite already established protocols and tools for emergency response, pathogen-related contamination events in the DWDN continue to affect communities. Communication gaps and the underutilization of available technologies play an important role in this.

Artificial intelligence shows a great potential for addressing these challenges, specifically by enabling rapid information extraction at the early stages of an emergency. Additionally, real-time modeling tools show good promise in better predicting pathogen fate, transport, health impact, and in supporting event management (e.g., source tracking, selection of sampling locations, and mitigation options).

Combining the benefits of recent developments in AI information/data extraction, and modeling tools to support decisions on how to minimize the negative impact of a contamination event could lead to fewer casualties and improve response in terms of operation, time, and coordination. It could also lead to reduced financial costs, better informing the public (improving trust), and better allocation of resources.

The activities of this research were part of the EU-funded PathoCERT (Pathogen Contamination Emergency Response Technologies). This project aimed to enhance the situational awareness of first (and second) responders by developing tools to support decision-making during pathogen contamination events.

1.2 Background

1.2.1 Artificial Intelligence and information extraction processes

Throughout the years, there have been many definitions of Artificial Intelligence (AI). In 1955, the Dartmouth Research Project first described the concept of AI as an effort of "*making a machine behave in ways that would be called intelligent if a human were so behaving*" (McCarthy et al., 1955). Nearly a decade later, scientist Marvin Minsky defined AI as "*the science of making machines do things that would require intelligence if done by men*" (Minsky, 1968). Given that both definitions are clear but somewhat generic, a more recent and elaborate definition focuses more on the practical aspect of AI, defining it as a "system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019).

AI has been applied in many fields and disciplines, including finance[3], education[4], computer science (Russell & Norvig, 2002) and biomedical sciences. AI algorithms have been developed for healthcare practices such as disease diagnosis and imaging (Dilsizian & Siegel, 2014), drug interactions (Christopoulou et al., 2020), radiology (Jha & Topol, 2016), and Electronic Health Records (EHR) (Mehta, & Devarakonda, 2018). As the volume of healthcare data grows, it has become increasingly difficult for healthcare practitioners and researchers to keep up. This is where AI applications became essential, as algorithms can "learn" features and identify patterns from large volumes of documents using Text Mining (TM) techniques and three basic subsets of AI, Machine Learning (ML), Deep Learning (DL) and Natural Language Processing (NLP).

1.2.1.1 Text mining

Text mining is a multidisciplinary scientific field that intersects with ML, statistics, NLP and linguistics. It is described as the process of extracting important patterns from textual data to gain knowledge and information. TM is often coupled with NLP (sometimes using ML techniques) for processing and analyzing textual information (Talib et al., 2016). A variety of fields associated with TM are displayed in Figure 1.3.

Particularly in the healthcare sector, where vast amounts of EHR and biomedical literature are available, text mining is extensively applied and is known as biomedical text mining. This specialization focuses on the application of text mining techniques to textual data and medical literature within biomedical domains, aiming to extract relevant information (Cohen, 2005). However, conventional text mining techniques that rely on large, annotated corpora[5] are typically unsuitable for biomedical text mining, as they do not inherently contain the specialized biomedical terminology needed by practitioners. The development of specialized corpora was therefore considered necessary to identify features and patterns from biomedical text (Ohno-Machado et al., 2013). For this purpose, the National Library of Medicine's Unified Medical Language System (UMLS) is the most common domain that provides controlled vocabularies and ontologies such as Medical Subject Headings (MeSH) (Lipscomb, 2000; Bodenreider, 2004).

---

[3] https://www.investopedia.com/terms/a/algorithmictrading.asp

[4] https://elearningindustry.com/artificial-intelligence-in-the-classroom-role

[5] In linguistics, a structured set of texts (which can nowadays be stored and processed electronically) is named corpus (plural corpora). Source: Wikipedia.

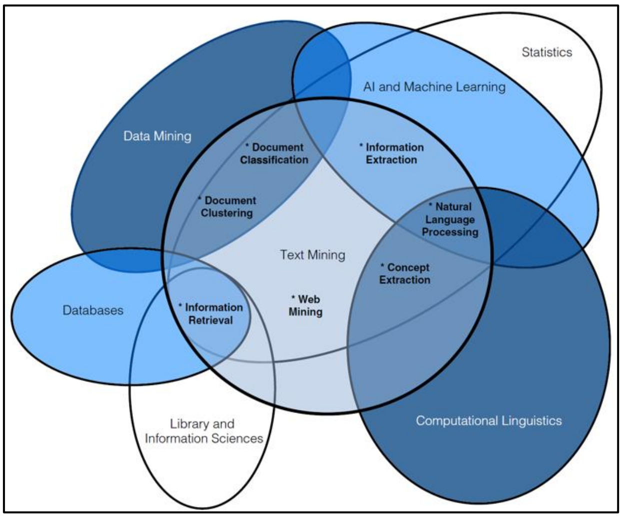


Figure 1.3. Venn diagram of Text mining and its interconnections with other fields. Source: Talib et al., 2016.

1.2.1.2 Machine learning

While AI broadly aims to mimic human behavior, ML is the scientific field that uses algorithms and statistical models to perform specific tasks by recognizing patterns, classifying data, and drawing conclusions from datasets without explicit human instructions. ML is a dynamic process that builds mathematical models using sample data. This can be either training[6] or testing[7] data and based on these, predictions can be made without requiring additional human interventions (Bishop, 2006). These models utilize supervised and unsupervised learning techniques. Supervised learning techniques apply functions that use training data to map a set of inputs to labeled outputs. In contrast, unsupervised learning techniques, as the name suggests, do not use pre-labeled outputs. Instead, these techniques require the algorithm to infer the structure of the data on its own, with minimal to no human supervision (Kaplan & Haenlein, 2019).

Many researchers have successfully used ML techniques to extract information related to study characteristics, such as disease-drug associations, from EHR, clinical studies, and Randomized Control Trials (RCT) (Chen et al., 2008; Uzuner et al., 2010; Kang & Weng, 2019). For example, Kiritchenko et al., 2010, developed ExaCT, an IE system that extracts 21 key trial characteristics from publications and helps curators review (using a user interface) and collect information from RCT. Their approach combined ML techniques involving both supervised and unsupervised methods, utilizing a Support Vector Machine (SVM) model for sentence classification along with rule-based techniques to extract exact values from text segments. Similarly, Patrick & Li, 2010, employed a multistage ML-based method using two different statistical classifiers -SVM and Conditional Random Fields (CRF) -along with rule-based methods, achieving nearly optimal results compared to other participants for automated extraction of medication information from clinical notes.

1.2.1.3 Deep learning

Deep learning is a subset of ML, under the broader umbrella of AI as Figure 1.4 suggests. It is characterized by models known as neural networks that mimic the structure and function of the human brain (LeCun et al., 2015). DL automates much of the feature extraction part of the data processing, which traditionally requires human intervention, allowing these models to learn high-level abstractions from data through multiple layers of processing. This means they can recognize and understand complex patterns in the data without being explicitly programmed to look for specific features. Typically, when DL is applied in text processing tasks, the model starts with individual characters or words, then learns to recognize phrases, and ultimately understands complex sentence structures, context, and semantics (Devlin et al., 2018).

---

[6] A set of examples used to fit the parameters of a model. The model is trained on this training dataset using a supervised learning method.
[7] A dataset used to provide an unbiased evaluation of a final model fit on the training dataset.

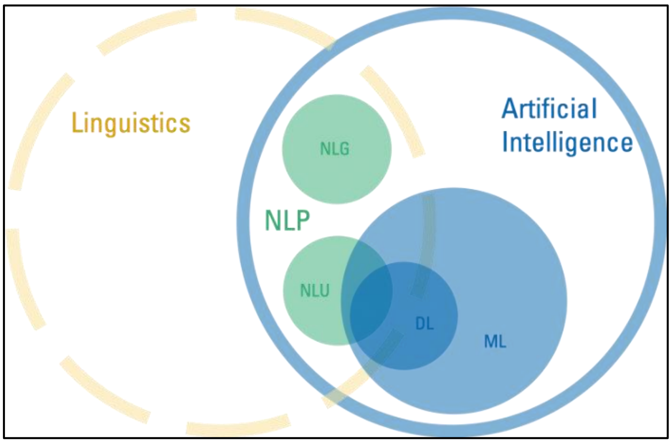


Figure 1.4. Interconnection between NLP and ML. NLP: Natural Language Processing, NLG: Natural Language Generation, NLU: Natural Language Understanding, ML: Machine Learning, DL: Deep Learning. Source: [*https://www.retresco.com/how-to-ai-natural-language-processing/*](https://www.retresco.com/how-to-ai-natural-language-processing/)

A significant architecture within DL for processing sequential data like text is the Transformer model, which uses self-attention mechanisms to dynamically weigh the importance of different inputs (Vaswani et al., 2017). This architecture is foundational in modern language processing tasks, including question-answering tasks (Rajpurkar et al., 2016). For example, DL techniques use models that can digest large volumes of text (structured or unstructured) and provide answers to questions (prompts) by understanding context and relevance.

Building on this over the last years, Large Language Models (LLMs), such as GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), represent an advanced step of the Transformer architecture (Vaswani et al., 2017; Devlin, 2018). LLMs are trained on -very- large datasets to generate coherent text based on the patterns they have learned. These models significantly enhance the capabilities of AI systems in tasks such as conversation, information extraction, and even generating entirely new content, demonstrating the practical applications of DL in natural language processing (Brown et al., 2020).

LLMs have shown remarkable capability in question-answering tasks, particularly when interacting with complex (unstructured) content often found in text documents such as PDFs. These models can read and interpret the content and then generate accurate answers to user queries. This functionality has been notably beneficial across various domains, including scientific research. For example, LLMs are used to extract and summarize findings from lengthy scientific papers, aiding researchers in quickly understanding study outcomes without reading the entire document (Lee et al., 2023, Liu et al., 2023). One of the main disadvantages of using LLMs for question-answering tasks is the tendency to hallucinate and generate fictitious responses (Yao et al., 2023).

1.2.1.4 Natural Language Processing

Natural Language Processing is another subset of AI focused on facilitating interactions between computers and humans, with the ultimate goal of enabling machines to understand, analyze, manipulate, and generate natural language. It combines linguistics and AI to facilitate machines to process natural human language. NLP is incorporated in several fields and applications including machine translation, user interfaces, speech recognition, ML, TM and broader AI disciplines (Chowdhury, 2003). Figure 1.4 illustrates the interconnectedness of NLP, ML, and DL, highlighting areas of overlap where ML/DL techniques are frequently employed in NLP tasks, particularly in data mining and information extraction.

1.2.2 State of the art in modeling substances in the drinking water network

EPANET, a widely used public domain software for hydraulic and water quality modeling, has improved the understanding of how (chemical) substances behave both in the water column and on pipe walls of DWDNs (Rossman, 2000). Over the last 25 years, it has been established as the *de facto* standard tool for water operators allowing them to simulate various “what-if” scenarios, and for researchers to develop and apply algorithms to tackle a plethora of research challenges in the DWDN.

Recognizing its limitations in modeling only single-species chemical dynamics, Shang et al. (2008b) developed an improved version known as EPANET-MSX (Multi Species Extension). This new tool allowed for the simulation of the reaction and transport dynamics of multiple interacting species, providing insights into the fate and transport of physical, chemical, and biological species.

Advancing the state of the art, Eliades et al. (2016) introduced the EPANET-MATLAB Toolkit, an open-source software that integrates EPANET (and EPANET-MSX) with MATLAB, a high-level programming environment. The toolkit allows the user to exploit MATLAB’s extensive computational and visualization capabilities while having access and enabling modifications to all the various aspects of a DWDN, such as executing direct calls to the EPANET library, running multi-species simulations, and creating or modifying networks.

1.2.2.1 Modeling pathogens and chlorine decay in the DWDN

Considerable research has been done on trying to model pathogen transport and chlorine dynamics in DWDNs, focusing on understanding microbial intrusion and chlorine decay mechanisms under various contamination scenarios.

Earlier work by Propato and Uber (2004) modeled a microbial intrusion incident in a chlorinated network. Their approach had simplifying assumptions such as that every node has an equal chance of being contaminated, and that intrusions were modeled as continuous. For the inactivation kinetics, the Chick-Watson equation was used:

$$\frac{dN}{dt} = -kC^{n}N \qquad 1$$

Where N is the number of organisms, k the inactivation rate constant, and C the disinfectant concentration raised to the power of n which represents the reaction order with respect to the disinfectant.

For the chlorine decay, simple first-order kinetics were used using the equation below:

$$\frac{\mathrm{dC}}{\mathrm{dt}} = -kC \qquad 2$$

Where k is the disinfectant decay rate constant.

In 2008, Betanzo et al. attempted to model the impact of microbial intrusion on chlorine levels within a DWDN by considering water quality parameters such as pH and temperature, using *Giardia* and *Escherichia coli 0157:H7* as the pathogens. They employed a simple first-order equation for chlorine decay without distinguishing between bulk and wall reactions. The inactivation kinetics followed again the Chick-Watson equation.

Helbling and VanBriesen (2009) tried to model the residual chlorine response to a microbial contamination event in a DWDN, using a parallel first-order model and incorporating a slow and fast chlorine reaction:

$$\frac{\mathrm{dC_{fast}}}{\mathrm{dt}} = -\mathrm{k_1 C_{fast}} \qquad 3$$

$$\frac{\mathrm{dC_{slow}}}{\mathrm{dt}} = -\mathrm{k_2 C_{slow}} \qquad 4$$

Where: C_{fast} is the fraction of chlorine that reacts rapidly with the contaminant ($mgCl - equiv * L^{-1}$), C_{slow} is the fraction of chlorine that reacts slowly with the contaminant ($mgCl - equiv * L^{-1}$), $k_1$and k_2 ($1/time$) are the reaction rates for the fraction of chlorine reacting rapidly and slowly respectively.

The microbial contaminants they used were *Escherichia coli, Staphylococcus epidermidis*, and *Mycobacterium aurum*. They concluded that the unique system architecture, the species-specific reaction kinetics, the demand patterns, as well as the initial chlorine concentration are factors that affect the chlorine demand signal propagation.

Further exploring the complexity of chlorine decay, Blokker et al. (2014a) studied the influence of stochastic water demands and temperature on residual chlorine modeling. They used a combined model from Vasconcelos et al. (2007) and Clark et al. (2011) with the assistance of EPANET-MSX which describes the chlorine decay as:

$$\frac{\mathrm{dC}}{\mathrm{dt}} = -\mathrm{k}_{20} * \exp\left(\frac{-\frac{\mathrm{E}}{\mathrm{R}} * (20 - \mathrm{T}_{\mathrm{water}})}{(273 + 20) * (273 + \mathrm{T}_{\mathrm{water}})}\right) * \mathrm{C} - \mathrm{k}_{\mathrm{r}} * \mathrm{C} - \mathrm{d} * \mathrm{k}_{\mathrm{mt}}^{\mathrm{e}} * \mathrm{C} \qquad 5$$

Where C is the concentration of free chlorine $(mgCl/L)$, K_{20} is the base value at a reference temperature of 20 °C, E/R is the activation coefficient (K), $Twater$ is the temperature of the drinking water in the DWDN (°C), Kr is the stagnant decay rate $(1/s)$, d, e are parameters in the net wall decay function, and k_{mt} is the coefficient of mass transfer to the pipe wall (m/s).

Their hypothesis was that a detailed hydraulic network model using stochastic water demands will exhibit different levels of chlorine residual than a conservative transport model that uses standard demand patterns. They concluded that the selection of the right chlorine decay model is more important than the stochastic demand model in calculating the chlorine residual levels. Another remark was that at certain locations of the DWDN, the inclusion of wall chlorine decay had a significant impact on the predicted chlorine residual. Finally, temperature can have a significant impact and should not be overlooked.

Pelekanos et al. (2021) used a parallel first-order model for both bulk and wall chlorine decay to evaluate the vulnerability of a water network to deliberate contamination attacks, focusing on nominal water demands. They demonstrated how the contamination location impacts the size of the exposed population.

Finally, Fisher et al. (2017) employed a two-reactant chlorine decay model that distinguished between fast and slow reacting agents, incorporating temperature because of its significant impact on bulk chlorine decay. The selected equations are given below:

$$\frac{\mathrm{dC_f}}{\mathrm{dt}} = -\,\mathrm{k_f C_f C} \qquad 6$$

$$\frac{\mathrm{dC_s}}{\mathrm{dt}} = -\,\mathrm{k_s C_s C} \qquad 7$$

$$\frac{\mathrm{dC}}{\mathrm{dt}} = \frac{\mathrm{dC_f}}{\mathrm{dt}} + \frac{\mathrm{dC_s}}{\mathrm{dt}} \qquad 8$$

Where C is the concentration of free chlorine $(mgC * L^{-1})$, C_f and C_s are the concentrations of fast and slow reducing agents $(mgC - equiv * L^{-1})$, k_f and k_s are fast and slow reaction rate coefficients $(LmgC^{-1}h^{-1})$.

1.2.2.2 Applications of risk assessment in DWDN modeling

Earlier work by Lieverloo et al. (2007) investigated the risk associated with fecal contamination in the DWDN. They calculated pathogen concentrations by using data from outbreak and non-outbreak contamination events, employing pathogen-to-coliform ratios. Those ratios were used to estimate fecal indicator concentrations and calculate the risk of infection. This approach highlighted the absence of hydraulic modeling to calculate and simulate more accurately pathogen concentrations in the DWDN, before assessing the risk of infection due to contamination.

Regarding the infection risk, besides calculating the number of contaminated nodes, or the total contaminant mass entering a node, a more detailed risk assessment that considers various factors (e.g. pathogen type, duration, and dose) exists, allowing for prediction of health risk with higher resolution (e.g. predict number of infections). This form of risk assessment is called Quantitative Microbial Risk Assessment (QMRA), and it has received attention over the last 2 decades. It has also been embedded in the WHO water-related guidelines (WHO, 2017). QMRA is a mathematical framework for evaluating infectious risks by combining scientific knowledge about pathogens (fate, transport, route of exposure, and health effects of human pathogens) with the effect of physical/mechanical barriers and mitigation actions (WHO, 2016). There are 4 steps associated with QMRA, summarized in Table 1.3.

The most common dose-response models (Step 3 from Table 1.3) used in QMRA are the Beta-Poisson model, its approximation, and the exponential model where each one has different characteristics, and it is suitable for specific cases. The Beta-Poisson model includes the hypergeometric function, and it is often used when there is variability in host susceptibility (Teunis et al., 2002). This model accounts for differences in immune response of each individual, which leads to more precise but also more computationally complex results. The approximation of Beta-Poisson is a simplified version of the hypergeometric model, retaining the core elements, while discarding the heavy computations. This model is usually chosen when the complexity of the hypergeometric function is not necessary, and it is suitable for low doses (Haas, 1983). Finally, the exponential model is the simplest since it assumes that there is uniform susceptibility among individuals (Haas, 1999). It is used when limited data is available, or when the host response to the pathogen is considered uniform.

The choice of the dose-response model depends on the pathogen type, the availability of data, and the level of precision required.

Table 1.3. Steps of QMRA. Adapted from WHO. Guidelines for Drinking-water Quality," 4th Edition, 2017.

Steps	Description
1. Hazard identification	Identify hazards associated with drinking water.
2. Exposure assessment	Determine the route, magnitude, and duration of exposure as well as the size and nature of population exposed.
3. Dose-response	Characterize the relationship between exposure and incidence of health effect with dose-response models.
4. Risk characterization	Integrate information from exposure and dose-response to estimate the risk.

Although most studies focus on applying QMRA to treated wastewater or contaminated bodies such as surface water (Owens et al., 2020), some research also examines the risk of microbial contaminants entering the DWDN. Microbial contaminants can enter the DWDN for multiple reasons including intermittent water supply (IWS), or intentional contamination. IWS happens when the delivery of water is not continuous, leading to periods of low (even no or negative) pressure in the pipes of a network, due to DWDN deficiencies, pipe breakages, or inadequate access to water (Bivins et al., 2017).

IWS can lead to pressure transients (low or negative-pressure periods allowing contamination ingress). The work from Besner et al. (2010), highlighted the implications of negative pressure events in a full-scale DWDN. The authors combined field pressure monitoring, hydraulic and transient analysis, and microbial characterization of external contamination sources. They used a leakage rate approach and orifice equations to estimate the contamination intrusion volumes, considering also factors such as external head pressure and orifice diameter.

The same group (Besner et al., 2011) reviewed microbial intrusions in DWDNs, aiming to understand the current methodologies for evaluating health risks. They discussed the use of transient analysis and hydraulic modeling (using EPANET) to simulate the transport of pathogens from entry points to consumer taps. The authors presented a conceptual QMRA model that includes causes of low and negative pressure events, pathways for contaminant entry, and the occurrence and transport of pathogens.

Although transient analysis in DWDNs can be considered as a specialized aspect of hydraulic modeling, modeling pressure transients can be combined also with conventional hydraulic modeling. For instance, Teunis et al. (2010) used both conventional hydraulic modeling with EPANET-MSX and transient analysis with a surge model (InfoSurge) to assess the risk from a virus intrusion in a DWDN caused by negative pressure transients using a hydraulic model and Monte Carlo simulations for random entry and dilution of the virus. They included inactivation kinetics and chlorine decay in their model, taking into account specific values for pH and temperature, while using a beta probability distribution for the dose-response. They also considered the coincidence of virus presence with the use of tap water. Their findings suggested that the likelihood of a consumer drinking contaminated water is low and that the spatial distribution of infection risk is largely uniform. As expected, the presence of chlorine residual in the DWDN reduced the infection risk.

Another effort described by Yang et al. (2011) involved surge modeling and hydraulic simulations to model a virus intrusion in a DWDN again due to pressure transients. The authors employed EPANET-MSX to integrate a Chick-Watson model that accounted for the inactivation kinetics of selected pathogens and chlorine decay, while their dose-response model was characterized by a beta-probability distribution of infectivity for a single norovirus particle. They concluded that the factors influencing the risk of viral infection were the duration of the negative pressure event and the number of affected nodes, without incorporating stochastic water demand or other water quality parameters.

Blokker et al. (2014b, 2018) developed a QMRA model to assess the risk from contamination events following main repairs in the DWDN, including the coincidence of opening the tap as the contaminant passes. They modeled the presence of pathogens like *Giardia*, *Campylobacter*, *Cryptosporidium*, and enterovirus, using the hypergeometric function distribution for the calculation of the dose-response, while applying stochastic water demands in the network. Their findings highlighted that the contamination concentration is the primary factor determining the ingested dose, while the chosen pathogen dose-response relationship significantly influences the infection risk. They did not account for chlorine decay modeling since usually no chlorine residual is applied in the Netherlands.

Schijven et al. (2016), evaluated exposure scenarios of intentional microbiological contamination in an unchlorinated DWDN using EPANET-MSX. Specifically, they investigated the effects of duration, concentration, exposure pathway, and pathogen infectivity on exposure and infection risk using the exponential dose-response model. They concluded that if the pathogen concentration is 10^6pathogens/L or more, the infection risk per event is close to 1, while a longer duration of an event leads to increased probability of exposure.

Vinas et al. (2022) developed a framework for estimating the risk of infection from cross-connection and backflow events in DWDNs. Their reference pathogens were *Campylobacter*, norovirus, and *Cryptosporidium*, using the Beta-Poisson dose-response model, including the hypergeometric function. For modeling the pathogen transport they used EPANET, while nominal water demands were applied. They applied QMRA integrating a Fault Tree Analysis approach for the estimation of contamination probabilities and infection risks in DWDNs. They found that the daily infection risk in Swedish DWDNs often exceeded acceptable levels, highlighting the importance of using local data for accurate assessments, while showcasing the usefulness of combining QMRA with the Fault Tree Analysis approach.

The same group, this time led by Odhiambo et al. (2023), estimated infection risks from pipe breaks and IWS in DWDNs. Once again, the Beta-Poisson distribution (hypergeometric function) for the calculation of dose-response was used, as well as nominal water demands, combining hydraulic modeling (EPANET) with QMRA to evaluate the intrusion of the same pathogens in a Swedish DWDN. They used multiple contamination scenarios using pathogen concentrations from field studies to assess intrusion volumes and pathogen entry in the DWDN. Their conclusion was that the pathogen concentration and the duration of contamination are the most influential factors of infection risk.

Shakibi (2022) applied QMRA to microbial intrusions during low-pressure events in DWDNs. Once again, the pathogens *Campylobacter*, norovirus, and *Cryptosporidium* were used, applying hydraulic modeling to simulate multiple scenarios (e.g. pump shutdowns, and pipe repairs) using nominal water demands in the network. For the QMRA, the approximation Beta-Poisson dose-response model was used to estimate the infection probabilities. The author concluded that the contamination duration and the location of the low-pressure events highly influence the volume of intrusion, while pump shutdowns pose a greater risk than pipe repairs.

Hatam et al. (2019) combined QMRA with pressure-driven hydraulic analysis (EPANET) to assess the health impact of accidental intrusion (low-pressure conditions) in a full-scale DWDN. The authors modeled various concentrations of *Cryptosporidium* from raw sewage under multiple intrusion scenarios of different durations using the hypergeometric function distribution for the calculation of the dose-response. They performed Monte Carlo simulations to account for customer water consumption behavior (coincidence of passage of contaminants at the tap). Their results suggested that the infection risk increases as soon as the duration and concentration increase, highlighting the need for rapid responses.

1.2.2.3 Overview of studies focusing on modeling inactivation kinetics, chlorine, and QMRA

Literature suggests that many researchers have used EPANET-MSX to model pathogen inactivation kinetics and chlorine decay in a DWDN, while hydraulic modeling has also been combined with QMRA to estimate infection risks under different conditions. Table 1.4 provides an overview of studies that have focused on modeling pathogens, chlorine decay, and applying QMRA in DWDNs, all using different combinations of relevant aspects.

Table 1.4. An overview of studies that focused on modeling pathogens, chlorine decay, and QMRA in the DWDN. P=Pathogen modeling, C= Chlorine decay modeling, Q= QMRA modeling.

	Study	**Pathogens/ indicators**	**Chlorine decay kinetics**	**Bulk chlorine decay**	**Wall chlorine decay**	**Temperature °C (min/max)**	**Water demand**	**Contamination type**	**TOC modeling**	**Dose-Response**	**Hydraulic modeling**
P,C	Propato and Uber, 2004	*Giardia*	First-order	Yes	No	5/20	Nominal	Microbial intrusion	No	N/A	EPANET-MSX
	Betanzo et al., 2008	*Giardia*, Escherichia coli 0157:H7	First-order	Yes	No	0.5/25	Nominal	Sewage intrusion	No	N/A	EPANET
	Helbling and VanBriesen, 2009	Escherichia Coli, *Staphylococcus epidermidis*, *Mycobacterium aurum*	Parallel first-order	Yes	Yes	N/A	Nominal	Microbial intrusion	No	N/A	EPANET-MSX
	Pelekanos et al., 2021	Bacterial concentration	Parallel first-order	Yes	Yes	14.6	Nominal	Intentional contamination	Yes	N/A	EPANET-MSX
C	Blokker et al., 2014a	N/A	First-order	Yes	Yes	Changes through time	Stochastic	Microbial contamination after main repairs	No	N/A	EPANET-MSX
	Fisher et al., 2017	N/A	Parallel second-order	Yes	Yes	20	Nominal	N/A	No	N/A	EPANET-MSX
P,C,Q	Teunis et al., 2010	Rotavirus, norovirus	First-order	Yes	No	10	Stochastic	Pressure transients	No	Beta	EPANET-MSX
	Yang et al., 2011	Norovirus	First-order	Yes	No	N/A	Stochastic	Pressure transients	No	Beta	EPANET-MSX
P,Q	Blokker et al., 2014b, 2018	*Giardia*, *Campylobacter*, *Cryptosporidium*, enterovirus	No	No	No	N/A	Stochastic	Microbial contamination after main repairs	No	Hyp/metric function	EPANET
	Schijven et al., 2016	Generic pathogenic microorganisms	No	No	No	N/A	Stochastic	Intentional contamination	No	Exponential	EPANET-MSX
	Vinas et al., 2022	*Campylobacter*, norovirus, *Cryptosporidium*	No	No	No	No	Nominal	Cross-connection	No	Beta-Poisson	EPANET
	Odhiambo et al., 2023	*Campylobacter*, norovirus, *Cryptosporidium*	No	No	No	No	Nominal	Pipe break, IWS	No	Beta-Poisson	EPANET
	Shakibi et al., 2022	*Campylobacter*, norovirus, *Cryptosporidium*	No	No	No	No	Nominal	Pump shutdowns, pipe repairs	No	Beta-Poisson	EPANET
	Hatam et al., 2019	*Cryptosporidium*	No	No	No	No	Stochastic	Accidental intrusion	No	Hyp/metric function	EPANET

1.2.3 Finding the contamination source

Over the last 20 years, there has been an increased effort on using data from ad hoc and/or routine sampling or sensor signals at nodes within a DWDN, to identify both the source and the time of a contamination (Adedoja et al., 2018). Often, scientists and experts try to model a -conservative-chemical rather than a microbial contaminant (Shang et al., 2002; Propato et al., 2010; Yang et al., 2014). The contamination source identification problem (CSI) is mainly considered a deterministic inverse problem where hydraulic calculations are used to track the path of water parcels to find the contamination source (Jerez et al., 2021). It is considered a challenging problem due to the computational burden associated with hydraulic calculations, and the non-uniqueness of the solutions in identifying the source. The 4 main reasons for non-unique solutions are:

1. The multiple paths water can take to reach a set of nodes (sampling) or sensors.
2. The absence of specific water quality sensors, which results in a lack of essential information about the reaction dynamics and the magnitude of the contamination.
3. The possibility of multiple sources (nodes) of contamination within the network.
4. The start and duration of contamination.

Therefore, any approach on CSI should be able to identify a potential area (multiple upstream contamination nodes), taking into account the associated inherent uncertainties on hydraulics, valve settings uncertainty, reaction dynamics (in case of chlorination), as well as false alarms triggered by nonspecific water quality sensors (Yang and Boccelli, 2014; Jerez et al., 2021).

1.2.3.1 Particle Backtracking

The first attempts on CSI were focused on the particle backtracking approach. Shang et al (2002) used a particle backtracking algorithm based on a Lagrangian[8] model where the contaminants were considered as particles that run in reverse time from the detection node to the source of contamination.

De Sanctis et al (2010), used a modified particle backtracking algorithm to identify the source of contamination by using binary sensor information. Although their results were promising, the model assumed perfect sensor accuracy and disregarded hydraulic uncertainties, which are critical in real-world applications.

Another simplified version of the particle backtracking approach, described by Eliades and Polycarpou (2012), computed pairs of nodes and times indicating where and when the contamination may have occurred. Their methodology was based on decision trees, expressing conditional statements such as *if-then-else* rules, to return a sequence of nodes for manual sampling.

1.2.3.2 Simulation-Optimization techniques

Simulation-optimization techniques[9] have also been broadly used to deal with the CSI problem. Laird et al (2005) developed an origin tracking algorithm that applied nonlinear, infinite-dimensional optimization, subject to differential constraints. Their method, tested on a small network of 469 nodes, highlighted the scalability challenges inherent in handling up to 210,000 variables, suggesting that such complexity could hinder real-time application in larger networks due to the computational burden.

Preis and Ostfeld (2008), combined the hydraulic and water quality software EPANET with a genetic algorithm (GA) to solve the CSI problem. Their objective function was to minimize the difference between measured and calculated concentration, and although they showed promising results, it required excessive computation using parallel computing. Yan et al (2017) attempted to mitigate the computational burden by using a cultural algorithm but yet again, there were uncertainties with respect to water demand and computational resources.

---

[8] The process of tracking particles through the integration of motion equations in given flow fields, providing insights into particle dispersion and turbulent flow dynamics (Ouellette and Bodenschatz, 2006).

[9] “The process of finding the best input variable values from among all possibilities without explicitly evaluating each possibility. The objective of simulation optimization is to minimize the resources spent while maximizing the information obtained in a simulation experiment” (Carson and Maria, 1997).

Despite significant advancements, a consensus on the optimal CSI methodology remains unclear (Adedoja et al., 2018). The field continues to face challenges related to computational demands, especially for real-time applications in extensive network settings. As noted by Rasekh and Brumbelow (2021), while simulation-optimization techniques offer a more sophisticated approach, they still face substantial obstacles due to the intensive computational time required for hydraulic calculations.

1.2.4 Network operational intervention

Water utilities are responsible for providing safe drinking water even when the integrity of their system is compromised. For this reason, a lot of research has been made regarding network operational interventions to better monitor and control, and restore DWDNs, and ensure that clean and safe water reaches consumers' taps.

1.2.4.1 Valve manipulation

Regarding network operational interventions in the DWDN, the most common is that of valve manipulation (opening or closing of valves) and it is used for various purposes. For example, Poulin et al (2008) proposed a heuristic approach to define isolation strategies, and thus identifying the valves to be closed by response teams in the field to tackle a drinking water contamination in the network. Although they presented promising results in two real-world DWDNs, their approach was based on a set of simplifying assumptions and needed to be further validated with other networks.

Another example of valve manipulation is presented in Abraham et al (2018) in an attempt to maximize the self-cleaning capacity of a DWDN by controlling the diurnal peak flow velocities in the pipes. The authors proposed an algorithm that enabled favorable changes in the flow velocities and thus maximized the self-cleaning capacity by identifying a set of isolating links to form a more branched network.

Mahmoud et al (2018) presented a methodology for the isolation of failure events (pipe burst or equipment failure) followed by operational interventions. It included a multiobjective optimization with the aim to minimize the negative impact on the customers and the corresponding number of interventions (a surrogate for operational costs). The results obtained (after applying the methodology in a real-world DWDN), demonstrated the effectiveness of the approach by highlighting Pareto optimal intervention strategies.

Vrachimis et al (2020) proposed an active contamination detection scheme by manipulating valves to drive flows from specific parts to designated nodes within a DWDN, thus enabling sensors to monitor water quality during contamination. The objective function in this study was to minimize the impact to the population by detecting the contaminant as fast as possible.

Finally, Moghaddam et al (2022) used a Particle Swarm Optimization (PSO) algorithm for the strategic closure of pipes within a DWDN to mitigate a (chemical) contamination event. Their approach included both pipe closures and hydrant activations and a combination of those two, while their objective function was to minimize the number of contaminated nodes.

1.3 Identified knowledge gaps

The literature review has shown that despite considerable research conducted by scholars on AI-based information extraction, modeling pathogen propagation and chlorine decay in a network, identifying the contamination source, manipulating valves, and assessing the risk of infection under the framework of QMRA, significant gaps remain that present opportunities for further research.

1.3.1 1st knowledge gap

1.3.1.1 AI-based information extraction

Although AI-based information extraction is becoming an established method in the biomedical sciences, the creation of similar applications to extract information about pathogen characteristics from the scientific literature is still largely unexplored. To streamline the process of acquiring fast information at the early stages of a pathogen contamination event, an automated approach could be highly beneficial. AI could rapidly extract and collate necessary information from the growing volume of scientific publications, but accuracy of the extracted information is key to support effective decision-making.

1.3.1.2 Evaluating LLMs

Besides relying on the traditional and efficient methods of ML and DL techniques to extract information on pathogen characteristics, a paradigm shift has emerged with the use of LLMs. Although many newly developed LLMs have shown remarkable capabilities, they often face difficulties when used for question-answering tasks in scientific publications due to their limitations in understanding context and adaptability. These models can struggle to grasp the detailed and specialized context of scientific texts (in our case related to environmental microbiology literature), leading to responses that may seem correct but are actually irrelevant or incorrect (hallucination) (Yao et al., 2023). Additionally, LLMs might not effectively adapt to new scientific knowledge that emerges after their training data was compiled, which could result in outdated or incomplete answers. Moreover, their application in answering questions related to historical information on contamination events has not been explored. Evaluating the performance of these models is crucial to ensure they can accurately handle complex scientific information and remain reliable and useful for scientific applications, especially considering they are increasingly viewed as the most effective approach for current and future information extraction/question-answering tasks.

1.3.2 2nd knowledge gap

Regarding modeling efforts in the DWDN, the literature review shows that there are still limitations in the way existing studies address pathogen contamination events. While different combinations of relevant aspects, such as pathogen propagation, bulk chlorine decay, and infection risk assessment, have been explored (as Table 1.4 suggest), critical elements like hydraulic uncertainty, stochastic water demands, and wall chlorine decay are often overlooked or inadequately addressed. The infection risk is typically calculated using relatively small pathogen concentrations (pressure transients). Furthermore, these aspects are rarely integrated comprehensively in a single study. Practical challenges, such as computational burdens and limited real-world applicability of existing approaches, further complicate the situation. The following sections focus on specific areas that require further attention.

1.3.2.1 Pathogen and chlorine modeling

Although many researchers have modeled the interaction of chlorine with pathogens in chlorinated DWDNs, not all the important modeling parameters have been considered together. With respect to hydraulics, most studies have been performed under nominal water demands in the DWDN. Additionally, the inactivation kinetics and chlorine decay have been described only under specific water quality parameters. Some of the studies were based on assumptions such as continuous pathogen intrusion, no wall chlorine decay, and no total organic carbon (TOC) degradation, using simplifications like first-order chlorine decay equations. Therefore, there is a need for an in-depth and comprehensive representation of the different reactions that take place in a DWDN, including chlorine decay and TOC degradation, as well as the associated hydraulics and pathogen transport. This should include integrating all current relevant knowledge on the transport and fate of pathogens, bulk and wall chlorine decay, fast and slow reactions of chlorine with TOC, TOC degradation, temperature variations and their influence on inactivation rate, and various pathogen inactivation kinetics using stochastic water demands instead of nominal.

1.3.2.2 QMRA and hydraulic modeling

The literature suggests that few studies have applied QMRA while considering all of the inherent complexity of hydraulic modeling, pathogen and chlorine dynamics. The infection risk of wastewater contamination (high pathogen and TOC concentrations) has not been covered sufficiently in the literature since most studies focus on the risks associated with pressure transients in the DWDN. The majority of them have used first-order chlorine decay focusing mainly on the bulk phase of pipes without considering the additional chlorine demand of wall phase and organic compounds, while for the calculation of dose-response, the approximation Beta-Poisson model has been mainly used. It is expected that using the hypergeometric function, along with incorporating stochastic water demands and isolating the tap water use while using a daily consumption distribution per person to calculate exposure to a wastewater contamination event, will provide a higher resolution. Additionally, a realistic representation of inactivation kinetics and chlorine decay (both bulk and wall, both fast and slow), and the consideration of TOC degradation will influence the fate and transport of pathogen concentration. This, in turn, could affect the exposure assessment and dose-response steps of QMRA.

1.3.2.3 Source identification during contamination

Although various approaches address the CSI problem, the inherent computational burden of hydraulic calculations remains a critical drawback, especially considering the need for immediate response during emergency events. Even in studies where computational time is minimized and authors have delivered promising results, these approaches are based on many assumptions, including the exclusion of hydraulic uncertainties (e.g., valve settings uncertainty), simplifications of water demand, and small-scale networks. The applicability of these methods in real-time emergency situations is still questionable, as water utilities cannot rely on them for decision-making, particularly since they usually do not use real-time contaminant detection sensors for source identification. This is why water utilities still depend on their experience and intuition to determine potential contamination locations and use simple generic models to estimate the contamination area. Therefore, there is a need to address the inherent hydraulic uncertainties in the CSI and apply a modeling approach that doesn't require heavy computations, is not heavily dependent on valve settings uncertainty and uses efficient sampling to iteratively narrow down the potential contamination area(s).

1.3.2.4 Valve manipulation

There are many approaches to improving water quality and protecting customers by manipulating valves to isolate or direct contamination through the DWDN. Most studies in the literature use simplifying assumptions in small-scale networks and focus on evolutionary algorithms for their optimization problems. Their primary goal is to minimize the number of interventions- as a surrogate for operational costs-, the mass injection rate of chlorine, or to minimize the impact by redirecting the contamination to sensors for faster detection (and thus faster mitigation measures). However, the objective of minimizing infection risk, as calculated from QMRA, has not been sufficiently covered.

When managing a drinking water contamination event, water utilities usually do not rely on the approaches mentioned above. Instead, they proceed based on their intuition and knowledge of network characteristics to close valves. This approach can be prone to errors, which can have significant implications for health risks. If the wrong pipes are closed, the contamination plume could be directed into areas where contamination would not have occurred otherwise. Therefore, there is a need for a comparison between current (traditional) and model-based approaches regarding managing a real-time pathogen contamination event in the DWDN.

1.4 Research Objectives and Questions

This thesis has two main objectives, and it will generate knowledge and methodologies to bridge the two identified knowledge gaps.

1.4.1 First objective

At the early stages of a contamination event, water utilities could enhance their emergency response by adopting AI-based information extraction techniques instead of manual literature reviews to get access to information. The first objective of this thesis is to develop a method that accurately extracts specific information on pathogen characteristics from scientific publications using AI and to evaluate various LLMs for their effectiveness in extracting information from scientific publications about pathogen contamination events of DW through a question-answering task. By integrating knowledge of pathogen characteristics and insights from similar past contamination events, including potential impacts and mitigation measures, water utilities can quickly assess exposure risks and gain a better understanding at the early stages of contamination. Therefore, the research question of the first objective to bridge the 1st knowledge gap can be described as:

"Is it feasible to use an automated approach to extract accurate information on waterborne pathogens and pathogen contamination events from the literature?"

Additionally, the following sub questions are formulated to explore the potential of applying AI-based information extraction to pathogen characteristics from the environmental microbiology literature, and to evaluate the performance of LLMs in answering questions about pathogen contamination events of DW.

Sub-question 1

"Can we use an AI-based model to extract accurate waterborne pathogens-associated information from the scientific literature?"

Chapter 2 answers this question by developing a model that extracts both quantitative and qualitative information from scientific publications about the waterborne pathogen *Legionella* using Deep Learning and Natural Language Processing techniques. Seven information keywords about *Legionella* were selected as general, explicit, and reproducible (waterborne) pathogen characteristics of both a qualitative and a quantitative nature. Fifty peer-reviewed scientific publications about *Legionella* were manually selected from the search engine PubMed and used for the implementation of the model. Besides using the evaluation metrics of precision, recall and F-score, a human evaluation was performed to evaluate the performance of the model.

Sub-question 2

"How effectively do LLMs answer questions related to health risks and mitigation measures about drinking water contamination events, found in the scientific literature?"

Chapter 3 answers this question by systematically assessing the performance of multiple open-source LLMs specifically on a question-answering task related to drinking water pathogen contamination events. A corpus of 188 scientific articles was compiled and relevant information was extracted for manual annotation. Using 23 questions related to pathogen contamination events of DW, the LLM with the highest performance (using specific evaluation metrics) was selected. A manual evaluation of the top-performing model's answer was conducted.

1.4.2 Second Objective

During a pathogen (wastewater) contamination event in the DWDN, the primary goal of a water utility is to execute a minimal yet effective set of actions rapidly to mitigate the incident and restore normal operations in real-time. The second objective of this thesis focuses on the response actions of a water utility during an emergency. Specifically, it aims to develop modeling tools that realistically simulate pathogen (wastewater) contamination events in a DWDN, assess the health impact (QMRA), and compare current (traditional) and model-based approaches to find the source of contamination (using sampling results for guidance) and proceed to mitigation measures (valve manipulation). To achieve this, an open access water quality benchmark model is created. The benchmark model integrates current relevant knowledge regarding the transport and fate of pathogens in chlorinated systems, chlorine decay (bulk and wall) and its reaction with TOC, inherent hydraulic and valve configuration uncertainties, stochastic water demands as well as health impact calculations through QMRA. Therefore, there are two research questions to bridge the 2^{nd} knowledge gap:

Research question 1

"What is the health impact of a wastewater contamination in the DWDN considering factors such as chlorine disinfection, hydraulic uncertainty, initial pathogen concentration, pathogen inactivation rate, contamination duration, and contamination location?

Sub-question 1.1

"How does chlorine disinfection influence pathogen concentration and infection risk?"

Sub-question 1.2

"What are the differences in infection risk between a chlorinated and a non-chlorinated network?"

Sub-question 1.3

"What is the impact of hydraulic uncertainty on infection risk?"

Sub-question 1.4

"What is the impact of initial pathogen concentration, inactivation rate, and contamination duration on infection risk?"

Sub-question 1.5

"Does the contamination location in a DWDN play a role in the infection risk?"

Chapter 4 addresses all those questions by presenting BeWaRE (Benchmark for Water network and Risk Evaluation), an open-access benchmark testbed that enables water utilities to prepare and respond effectively to such contaminations. The testbed integrated all current relevant knowledge on pathogen transport, bulk and wall chlorine decay, fast and slow chlorine reactions with TOC, TOC degradation, stochastic water demands, and hydraulic uncertainty. The health impact was calculated using QMRA focusing on three waterborne pathogens: Enterovirus, *Campylobacter*, and *Cryptosporidium.* Synthetic household-level water demand time series were used to model the individual water consumption timing and calculate the infection risk (exposure via ingestion). The role of contamination location in infection risk was addressed. A nominal range sensitivity analysis was performed for the initial pathogen concentration, inactivation rate, and contamination duration.

Research question 2

"What is the added value of using modeling tools to support decision making during wastewater contamination events in the DWDN?"

Sub-question 2.1

"Can real-time pathogen sampling information be effectively utilized to identify the source of contamination in the DWDN?"

Sub-question 2.2

"What are the comparative advantages of applying model-based tools over traditional methods for handling wastewater contamination events in the DWDN?"

Sub-question 2.3

"How do evolutionary algorithms compare to expert judgment in determining the most suitable valve closure sequence for contamination mitigation?"

Sub-question 2.4

"To what extent does maintaining an up-to-date representation of the DWDN's valve configuration, impact the efficacy of finding the contamination source using modeling tools?"

Chapter 5 answers these questions by presenting a case study in the Netherlands as part of the EU-funded PathoCERT project. The objective was to systematically compare the efficacy of traditional and model-based decision-making in responding to wastewater contamination events in the DWDN when the source is unknown. The current approach on finding the contamination source was compared with the analytical tool named PathoINVEST (Pathogen contamination INVESTigation decision support system), a tool that integrates a QGIS plugin with the software EPANET-MATLAB Toolkit and the BeWaRE benchmark testbed. The comparison metrics regarding the source identification included the duration to find the source, the number of samples required, and the infection risk of the urban population at the time the source was identified. Regarding mitigation measures, the current (traditional) approach on closing valves based on expert judgement, was compared with one of the PathoINVEST features that uses evolutionary algorithms to find the optimal combination of closure valves. The problem of valve configuration uncertainty was also addressed.

References

Abraham, E., Blokker, M., & Stoianov, I. (2018). Decreasing the discoloration risk of drinking water distribution systems through optimized topological changes and optimal flow velocity control. *Journal of Water Resources Planning and Management*, *144*(2), 04017093.

Besner, M. C., Ebacher, G., Jung, B. S., Karney, B., Lavoie, J., Payment, P., & Prévost, M. (2010). Negative pressures in full-scale distribution system: Field investigation, modeling, estimation of intrusion volumes and risk for public health. *Drinking Water Engineering and Science*, *3*(2), 101-106.

Besner, M. C., Prévost, M., & Regli, S. (2011). Assessing the public health risk of microbial intrusion events in distribution systems: Conceptual model, available data, and challenges. *Water research*, *45*(3), 961-979.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

Bivins, A. W., Sumner, T., Kumpel, E., Howard, G., Cumming, O., Ross, I., ... & Brown, J. (2017). Estimating infection risks and the global burden of diarrheal disease attributable to intermittent water supply using QMRA. *Environmental science & technology*, *51*(13), 7542-7551.

Bjelkmar, P., Hansen, A., Schönning, C., Bergström, J., Löfdahl, M., Lebbad, M., ... & Lindh, J. (2017). Early outbreak detection by linking health advice line calls to water distribution areas retrospectively demonstrated in a large waterborne outbreak of cryptosporidiosis in Sweden. *BMC Public Health*, *17*, 1-10.

Blokker, M., Smeets, P., & Medema, G. (2014b). QMRA in the drinking water distribution system. *Procedia Engineering*, *89*, 151-159.

Blokker, M., Smeets, P., & Medema, G. (2018). Quantitative microbial risk assessment of repairs of the drinking water distribution system. *Microbial Risk Analysis*, *8*, 22-31.

Blokker, M., Vreeburg, J., & Speight, V. (2014a). Residual chlorine in the extremities of the drinking water distribution system: the influence of stochastic water demands. *Procedia Engineering*, *70*, 172-180.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, *32*(suppl_1), D267-D270.

Braeye, T., De Schrijver, K., Wollants, E., Van Ranst, M., & Verhaegen, J. (2015). A large community outbreak of gastroenteritis associated with consumption of drinking water contaminated by river water, Belgium, 2010. *Epidemiology & Infection*, *143*(4), 711-719.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Cann, K. F., Thomas, D. R., Salmon, R. L., Wyn-Jones, A. P., & Kay, D. (2013). Extreme water-related weather events and waterborne disease. *Epidemiology & Infection*, *141*(4), 671-686.

Carson, Y., & Maria, A. (1997, December). Simulation optimization: methods and applications. In *Proceedings of the 29th conference on Winter simulation* (pp. 118-126).

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. *Journal of the American Medical Informatics Association*, *15*(1), 87-98.

Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., & Ananiadou, S. (2020). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, *27*(1), 39-46.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

Craun, G. F., & Calderon, R. L. (2001). Waterborne disease outbreaks caused by distribution system deficiencies. *Journal-American Water Works Association*, *93*(9), 64-75.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports*, *16*(1), 441.

Eliades, D. G., & Polycarpou, M. M. (2012). Water contamination impact evaluation and source-area isolation using decision trees. *Journal of Water Resources Planning and Management*, *138*(5), 562-570.

Fewtrell, L., Kay, D., Watkins, J., Davies, C., & Francis, C. (2011). The microbiology of urban UK floodwaters and a quantitative microbial risk assessment of flooding and gastrointestinal illness. *Journal of flood risk management*, *4*(2), 77-87.

Fisher, I., Kastl, G., & Sathasivan, A. (2017). A comprehensive bulk chlorine decay model for simulating residuals in water distribution systems. *Urban Water Journal*, *14*(4), 361-368.

Geudens, P. J. J. G., & Grootveld, J. (2017). Dutch drinking water Statistics 2017. *VEWIN: Den Haag, The Netherlands*.

Giammanco, G. M., Bonura, F., Urone, N., Purpari, G., Cuccia, M., Pepe, A., ... & De Grazia, S. (2018). Waterborne Norovirus outbreak at a seaside resort likely originating from municipal water distribution system failure. *Epidemiology & Infection*, *146*(7), 879-887.

Haas, C. N. (1983). Estimation of risk due to low doses of microorganisms: a comparison of alternative methodologies. *American journal of epidemiology*, *118*(4), 573-582.

Haas, C. N., Rose, J. B., & Gerba, C. P. (1999). Quantitative microbial risk assessment.

Hatam, F., Blokker, M., Besner, M. C., Ebacher, G., & Prévost, M. (2019). Using nodal infection risks to guide interventions following accidental intrusion due to sustained low pressure events in a drinking water distribution system. *Water*, *11*(7), 1372.

Hrudey, S. E., & Hrudey, E. J. (2004). *Safe drinking water*. IWA publishing.

Jha, S., & Topol, E. J. (2016). Adapting to artificial intelligence: radiologists and pathologists as information specialists. *Jama*, *316*(22), 2353-2354.

Kang, T., Zou, S., & Weng, C. (2019). Pretraining to recognize PICO elements from randomized controlled trial literature. In *MEDINFO 2019: Health and Wellbeing e-Networks for All* (pp. 188-192). IOS Press.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15-25.

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

Kuhn, K. G., Falkenhorst, G., Emborg, H. D., Ceper, T., Torpdahl, M., Krogfelt, K. A., ... & Mølbak, K. (2017). Epidemiological and serological investigation of a waterborne Campylobacter jejuni outbreak in a Danish town. *Epidemiology & Infection*, *145*(4), 701-709.

Laine, J., Huovinen, E., Virtanen, M. J., Snellman, M., Lumio, J., Ruutu, P., ... & Pirkanmaa Waterborne Outbreak Study Group. (2011). An extensive gastroenteritis outbreak after drinking-water contamination by sewage effluent, Finland. *Epidemiology & Infection*, *139*(7), 1105-1113.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lee, Y., Lee, K., Park, S., Hwang, D., Kim, J., Lee, H. I., & Lee, M. (2023, July). QASA: advanced question answering on scientific articles. In *International Conference on Machine Learning* (pp. 19036-19052). PMLR.

Lieverloo, J. H. M., Blokker, E. M., & Medema, G. (2007). Quantitative microbial risk assessment of distributed drinking water using faecal indicator incidence and concentrations. *Journal of Water and Health*, *5*(S1), 131-149.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, *88*(3), 265.

Liu, J., Jin, J., Wang, Z., Cheng, J., Dou, Z., & Wen, J. R. (2023). Reta-llm: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*.

Mac Kenzie, W. R., Schell, W. L., Blair, K. A., Addiss, D. G., Peterson, D. E., Hoxie, N. J., ... & Davis, J. P. (1995). Massive outbreak of waterborne Cryptosporidium infection in Milwaukee, Wisconsin: recurrence of illness and risk of secondary transmission. *Clinical Infectious Diseases*, *21*(1), 57-62.

Mahmoud, H. A., Kapelan, Z., & Savić, D. (2018). Real-time operational response methodology for reducing failure impacts in water distribution systems. *Journal of Water Resources Planning and Management*, *144*(7), 04018029.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). A proposal for the dartmouth summer research project on artificial intelligence, august 1955. *URL: http://raysolomonoff. com/dartmouth/boxa/dart564props. pdf*.

Mehta, N., & Devarakonda, M. V. (2018). Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey?. *Journal of Allergy and Clinical Immunology*, *141*(6), 2019-2021.

Minsky, M. 1968, Semantic information processing. *Cambridge, Mass*.

Moghaddam, A., Afsharnia, M., Mokhtari, M., & Peirovi-Minaee, R. (2022). Management and health risk assessment of chemical contamination events in water distribution systems using PSO. *Environmental Monitoring and Assessment*, *194*(5), 362.

Odhiambo, M., Viñas, V., Sokolova, E., & Pettersson, T. J. (2023). Health risks due to intrusion into the drinking water distribution network: hydraulic modeling and quantitative microbial risk assessment. *Environmental Science: Water Research & Technology*, *9*(6), 1701-1716.

Ohno-Machado, L., Nadkarni, P., & Johnson, K. (2013). Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature. *Journal of the American Medical Informatics Association*, *20*(5), 805-805.

Ouellette, N. T., Xu, H., & Bodenschatz, E. (2006). A quantitative study of three-dimensional Lagrangian particle tracking algorithms. *Experiments in Fluids*, *40*, 301-313.

Owens, C. E., Angles, M. L., Cox, P. T., Byleveld, P. M., Osborne, N. J., & Rahman, M. B. (2020). Implementation of quantitative microbial risk assessment (QMRA) for public drinking water supplies: Systematic review. *Water research*, *174*, 115614.

Patrick, J., & Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, *17*(5), 524-527.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Risebro, H. L., Doria, M. F., Andersson, Y., Medema, G., Osborn, K., Schlosser, O., & Hunter, P. R. (2007). Fault tree analysis of the causes of waterborne outbreaks. *Journal of water and health*, *5*(S1), 1-18.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Schijven, J., Forêt, J. M., Chardon, J., Teunis, P., Bouwknegt, M., & Tangena, B. (2016). Evaluation of exposure scenarios on intentional microbiological contamination in a drinking water distribution network. *Water research*, *96*, 148-154.

Shakibi, M. (2022). Hydraulic Modeling and Quantitative Microbial Risk Assessment of Intrusion in Water Distribution Networks Under Sustained Low-Pressure Situations.

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414-418.

Tangena, B. (2018). Het gaat bijna altijd goed: Drinkwaterincidenten- statistieken, analyses en anekdotes. Het Boekenschap. https://www.hetboekenschap.nl/product/drinkwaterincidenten/ (ISBN: 9789083126609)

Teunis, P. F., Chappell, C. L., & Okhuysen, P. C. (2002). Cryptosporidium dose-response studies: variation between hosts. *Risk analysis*, *22*(3), 475-485.

Teunis, P. F. M., Xu, M., Fleming, K. K., Yang, J., Moe, C. L., & LeChevallier, M. W. (2010). Enteric virus infection risk from intrusion of sewage into a drinking water distribution network. *Environmental science & technology*, *44*(22), 8561-8566.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, *17*(5), 514-518.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Viñas, V., Sokolova, E., Malm, A., Bergstedt, O., & Pettersson, T. J. (2022). Cross-connections in drinking water distribution networks: Quantitative microbial risk assessment in combination with fault tree analysis and hydraulic modeling. *Science of the Total Environment*, *831*, 154874.

World Health Organization. (2011). *Small-scale water supplies in the pan-European region: background, challenges, improvements*. World Health Organization. Regional Office for Europe.

World Health Organization. (2016). Quantitative microbial risk assessment: application for water safety management.

World Health Organization. (2017). Guidelines for drinking-water quality: first addendum to the fourth edition.

Yang, J., LeChevallier, M. W., Teunis, P. F., & Xu, M. (2011). Managing risks from virus intrusion into water distribution systems due to pressure transients. *Journal of water and health*, *9*(2), 291-305.

Yao, J. Y., Ning, K. P., Liu, Z. H., Ning, M. N., & Yuan, L. (2023). Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*

Chapter 2

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

Published as original research article in the International Journal of Hygiene and Environmental Health, 245,114018. Paraskevopoulos, S., Smeets, P., Tian, X., & Medema, G. (2022).
https://doi.org/10.1016/j.ijheh.2022.114018

Chapter 2- Artificial Intelligence to extract information on pathogen characteristics from scientific publications

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.

Keywords: Artificial intelligence; Information extraction; Exposure assessment; Scientific publications; *Legionella*

2.1 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 & 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 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.

2.2 Materials and method

2.2.1 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 2.1).

Table 2.1. The desired extracted information (Information Keywords) from scientific publications regarding the waterborne pathogen Legionella. The 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

2.2.2 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 these 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.

2.2.3 Template filling

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 2.2 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, vaguer, and free text information.

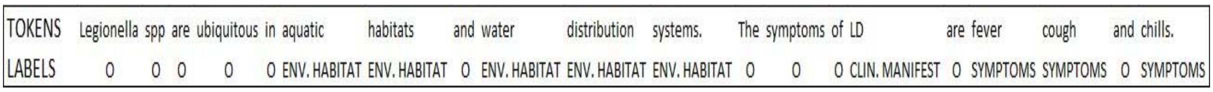
Table 2.2. Example of template filling extracting information from a scientific publication.

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

2.2.4 Information Extraction task

2.2.4.1 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 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”[10]. Figure 2.1 serves as an example of the labeling process.



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

Figure 2.1. 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”.

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.

2.2.5 Overall architecture

Fitting the overall architecture into a general workflow resulted in the following process (Figure 2.2).

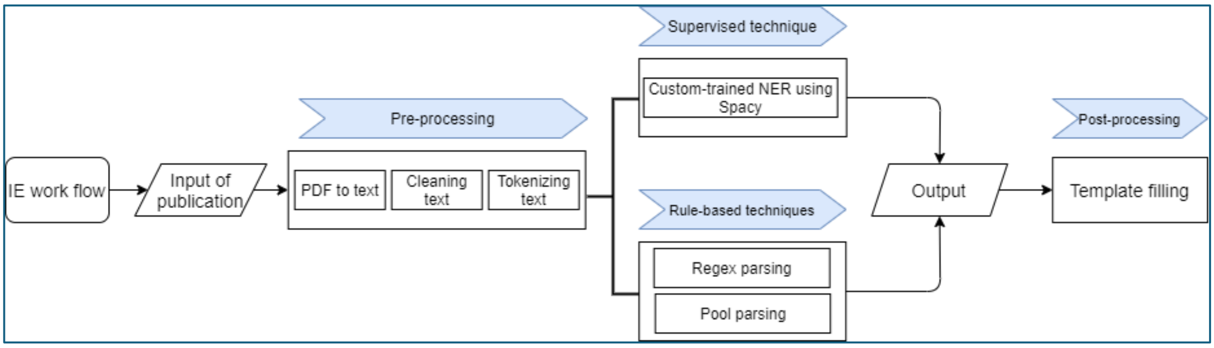


Figure 2.2. The workflow of the IE task starts with the input of publication. Next, the publication gets converted to text, cleaned, and tokenized as part of the pre-processing step. The next part includes supervised and rule-based techniques for the extraction of information. Finally, the output of this process gets filled in a template as part of the post-processing step.

---

[10] 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".

2.5.1 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 preprocessing 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[11] 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

---

[11] 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).

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).

2.2.6 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).

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

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

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

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).

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

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.

2.3 Results

2.3.1 Influence of the number of publications on evaluation metrics

Figure 2.3a shows that by increasing the number of publications, all metrics improved and the standard deviation of cross-validation regarding precision in Figure 2.3b 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.

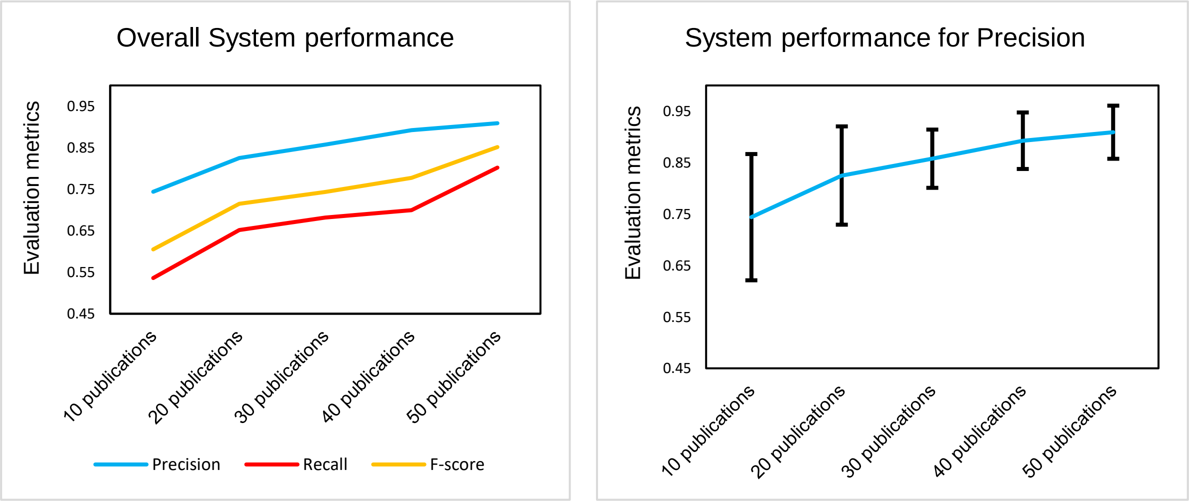


Figure 2.3. a) System performance under varying numbers of publications. b) System performance and standard deviation for precision under varying numbers of publications.

2.3.2 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 2.3 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.

Table 2.3. Confusion matrix of the custom-trained NER performance.

		Predicted labels				
	Information keywords	**Clin. Man/on**	**Env. habitat**	**O**	**Route of transmission**	**Source of Exposure**
Actual labels	**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**

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 2.4). 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 2.4. Confusion matrix of the rule-based techniques.

		Predicted labels			
	Information keywords	**Incubation period**	**O**	**Species**	**Symptoms**
Actual labels	**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 2.5 and 2.6 give an overview of the evaluation metrics of the system for the supervised and rule-based techniques. For the custom-trained NER in Table 2.5, 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 & 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.

Table 2.5. Classification report on 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	-

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

Table 2.6. Classification report on 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	-

2.3.3 Alternative evaluation with new publications

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

2.3.3.2 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 2.7 supports this argument. Although the sample is small and conclusions cannot be drawn, the evaluation metrics of both precision and recall are high. Table 2.8 depicts the extraction of information (and comparison) for 2 publications as an example. The rest of the comparison tables can be found in the supplementary material.

Table 2.7. 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

	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”	-	-

Table 2.8. Comparison between the system’s performance and manual extraction of IK from 2 publications. Red highlighted shade= erroneous results. Red bold font= Missed result (either by the IE model or by the manual extraction).

2.4 Discussion

2.4.1 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 Kiritcenko 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.

2.4.2 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.

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

2.4.4 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).

2.4.4.1 Extrapolate the process to other pathogens and/or fields

Although this Chapter 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.

2.5 Conclusions

This Chapter 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.

Supplementary material

Table S2. 1 Total number of scientific publications and the combination of search terms.

Search engine	**Search terms**	**Searched fields**	**Number of publications**	**New publications**
PubMed	(Legionella) AND (outbreak) AND (emergency)	All fields	140	140
	(Legionellosis) AND (outbreak) AND (emergency)	All fields	129	35
	(Legionella) AND (emergency)	All fields	391	251
	(Legionella) AND (incubation) AND (source of exposure)	All fields	10	9
	(Legionella) AND (incubation) AND (source of exposure) AND (transmission)	All fields	4	0
	(Legionella) AND (incubation) AND (transmission)	All fields	23	19
Total number of scientific publications				**454**

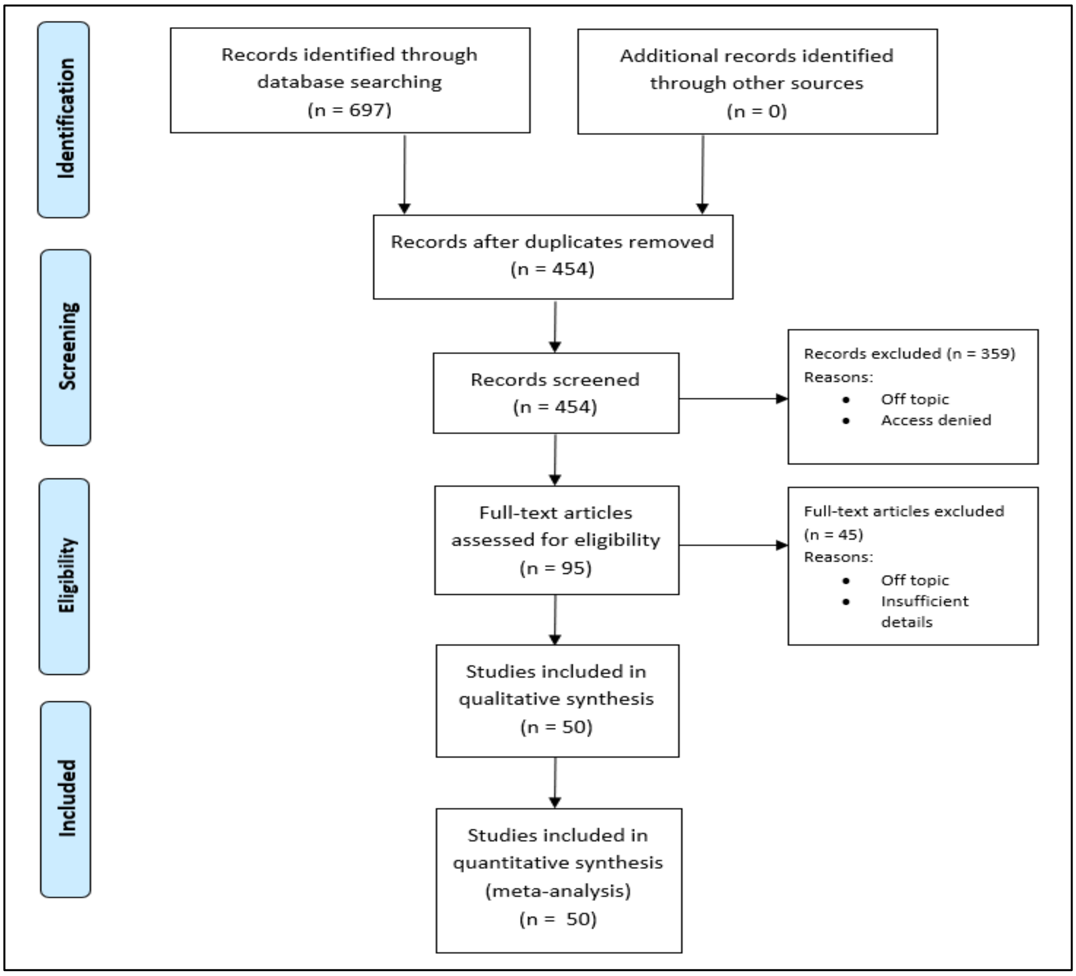


Figure S2. 1 Flow diagram on the selection of scientific publications adopted by the PRISMA guidelines.

Table S2. 2 List with the 50 selected scientific publications related to Legionella.

Title	**Reference**
A case of legionella pneumonia caused by home use of continuous positive airway pressure	(Schnirman et al., 2017)
A cluster of legionnaires' disease and associated pontiac fever morbidity in office workers, Dublin, June-July 2008	(Ward et al., 2010)
A community outbreak of Legionnaires' disease in Geneva, Switzerland, June to September 2017	(Zanella et al., 2018)
A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999	(Den Boer et al., 2002)
A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak	(Reuter et al., 2013)
Air-conditioner cooling towers as complex reservoirs and continuous source of Legionella pneumophila infection evidenced by a genomic analysis study in 2017, Switzerland	(Wüthrich et al., 2019)
An outbreak of travel-associated legionnaires disease and pontiac fever: The need for enhanced surveillance of travel-associated legionellosis in the United States	(Benin et al., 2002)
An unusually long-lasting outbreak of community-acquired Legionnaires' disease, 2005-2008, Italy	(Scaturro et al., 2014)
Antibiotic susceptibility of Legionella strains isolated from public water sources in Macau and Guangzhou	(Xiong et al., 2016)
Automobile windshield washer fluid: A potential source of transmission for Legionella	(Schwake et al., 2015)
Barrow-in-Furness: A large community legionellosis outbreak in the UK	(Bennett et al., 2014)
Outbreak Investigations and Identification of Legionella in Contaminated Water	(Lee, S., & Lee, 2013)
Community-acquired pneumonia: Role of atypical organisms	(Cosentini et al., 2001)
Confirmed and Potential Sources of Legionella Reviewed	(van Heijnsbergen et al., 2015)
Could it be Legionella?	(Darby and Buising, 2008)
Legionella pneumophila and Other Legionella Species Isolated	(Amemura-Maekawa et al., 2018)
Detection and identification of Legionella species from groundwaters	(Brooks et al., 2004)
Epidemiology and clinical management of Legionnaires' disease	(Phin et al., 2014)
Existence and control of Legionella bacteria in building water systems: A review	(Springston and Yocavitch, 2017)
Lessons From an Outbreak of Legionnaires’ Disease on a Hematology-Oncology Unit	(Francois Watkins et al., 2017)
Compost and Legionella longbeachae : an emerging infection?	(Currie and Beattie, 2015)
Legionella longbeachae detected in an industrial cooling tower linked to a legionellosis outbreak, New Zealand, 2015; possible waterborne transmission?	(Thornley et al., 2017)
Cutaneous Legionella longbeachae Infection in Immunosuppressed Woman, United Kingdom	(Harris and Battersby, 1990)
Legionella occurrence in municipal and industrial wastewater treatment plants and risks of reclaimed wastewater reuse: Review	(Caicedo et al., 2019)
Legionella pneumophila levels and sequence-type distribution in hospital hot water samples from faucets to connecting pipes	(Bédard et al., 2019)
Legionella: A reemerging pathogen	(Herwaldt and Marra, 2018)
Legionella: From environmental habitats to disease pathology, detection and control	(Atlas, 1999)
Legionellosis acquired through a dental unit: a case study	(Schönning et al., 2017)
Legionellosis on the rise: A review of guidelines for prevention in the United States	(Parr et al., 2015)
Legionellosis outbreak associated with asphalt paving machine, Spain, 2009	(Coscollá et al., 2010)
Legionnaires' disease	(Cunha et al., 2016)
Legionnaires' disease at a Dutch flower show: Prognostic factors and impact of therapy	(Lettinga et al., 2002)
Legionnaire's Disease in Compromised Hosts	(Lanternier et al., 2017)

Legionnaires' Disease in Hotels and Passenger Ships: A Systematic Review of Evidence, Sources, and Contributing Factors	(Mouchtouri and Rudge, 2015)
Legionnaire's Disease Since Philadelphia: Lessons Learned and Continued Progress	(Cunha and Cunha, 2017)
Legionnaires’ disease in Europe, 2011 to 2015	(Beauté, 2017)
Microbiological diagnosis and molecular typing of Legionella strains during an outbreak of legionellosis in Southern Germany	(Essig et al., 2016)
Microbiology and Epidemiology of Legionnaire's Disease	(Burillo et al., 2017)
Multiple sources of the outbreak of legionnaires’ disease in Genesee County, Michigan, in 2014 and 2015	(Smith et al., 2019)
New approach to environmental investigation of an explosive legionnaireś disease outbreak in Spain: Early identification of potential risk sources by rapid Legionella spp immunosensing technique	(Cebrián et al., 2018)
Outbreak detection and secondary prevention of Legionnaires' disease: A national approach	(Den Boer et al., 2007)
Outbreak of Legionnaires' disease associated with cooling towers at a California state prison, 2015	(Lucas et al., 2018)
Outbreak of Pontiac fever due to Legionella anisa	(Fenstersheib et al., 1990)
Outbreaks of Legionnaires’ Disease and Pontiac Fever 2006–2017	(Hamilton et al., 2018)
Results from the national Legionella outbreak detection program, the Netherlands, 2002–2012	(Den Boer et al., 2015)
Huhumidifier-associated paediatric legionnaires' disease, Israel, February 2012	(Moran-Gilad et al., 2012)
Legionella contamination in warm water systems: A species-level survey	(Dilger et al., 2018)
Legionella infection and control in occupational and environmental health	(Fujii and Yoshida, 1998)
Legionella longbeachae and legionellosis	(Whiley and Bentham, 2011)
Severe Pneumonia Caused by Legionella pneumophila	(Chahin and Opal, 2017)

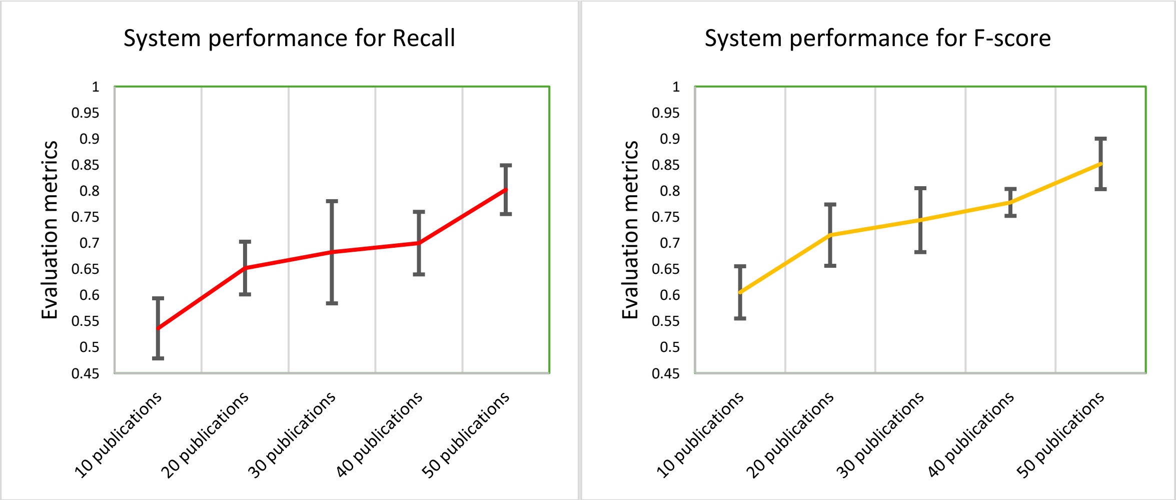


Figure S2. 2 System performance and standard deviation for Recall (left) and F-score (right) under different number of publications.

Table S2. 3 Comparison between the system's performance and manual extraction of IK from 10 new publications. Red highlighted shade= erroneous results. Red bold font= Missed result (either by the IE model or by the manual extraction).

	Epidemiology of Legionnaires Disease, Hong Kong, China 2005−2015 (Leung et al., 2020)		Evaluation of Legionella spp. Colonization in Residential Buildings Having Solar Thermal System for Hot Water Production (Totaro et		Identification of two aptamers binding to Legionella pneumophila with high affinity and specificity(Saad et al., 2020)	
	System	**Manual extraction**	**System**	**Manual extraction**	**System**	**Manual extraction**
Incubatio	-	-	-	-	-	-
Species	“ Legionella pneumophila”	“ Legionella pneumophila”	“ Legionella pneumophila”	“ Legionella	“ Legionella pneumophila”	“ Legionella
Symptoms	“ gastrointestinal” , ” cough” , “ fever”	“ fever” , “ couch” , “ **shortness of breath**” , “ gastrointestinal”	-		-	-
Cl. Manif/on	“ variables” , “ affiliation” , “ severe disease” , “ legionellosis” , “ acquired pneumonia” , “ Legionnaires disease” , “ rhabdomyolysis” , “ renal”	“ Legionnaires disease” , “ pneumonia” , “ renal function” , **“ septic shock”** , “ rhabdomyolysis”	“ Legionnaires disease” , **“ legionellosis”**	“ Legionnaires disease”	“ **pneumonia** legionnaires” , “ legionnaires disease ld” ,” legionellosis”	“ Legionnaires disease” , **“ Pontiac fever”** , “ legionellosis”
Source of exposure	water taps” , shower hose” , “ demonstration” , “ samples” , “ showers” , “ car air conditioning systems” , “ cooling towers whirlpools decorative fountains humidifiers” , “ aerosolized washer fluid” , “ cooling towers” , air conditioning system cabin air filter” , included septic” , “ water filters” , “ car cabin air” , “ respiratory equipment” , “ windshield washer fluid” , “ prac” , “ households” , “ respiratory failure”	“ Cooling towers” , “ whirlpools” , “ decorative fountains” . “ humidifiers” , “ respiratory equipment” , “ water tap” , “ shower head” , “ shower hose” , “ water filters, “ air conditioning system” , “ cabin air filter” , “ windshield washer fluid” , **“ residential potable water systems”**	“ cold water” , “ storage tanks included” , “ pipelines application” , “ domestic water plant” , “ tank receives” , “ of cold” , “ systems” , “ thermal systems” , “ colonized cold water” , “ aerosols” , “ **power plants”** , “ aqueduct municipal” , **“ pipelines”** , “ water system” , “ samples” , “ warm water production”	“ municipal water” , “ water storage tank” , “ hot and cold water” , “ aerosols” , “ **water plumbing systems**”	“ select aptamers” , “ cells” , “ detects dead cells” , “ alone without cells” , “ positive colonies” , “ monocytogenes” , “ trols included” , “ without cells” , “ aerogenes” , “ aptamers pools” , “ cooling tower” , “ tower” , “ ambion” , “ r8” , “ vbnc cells” , “ r10c1 aptamers” , “ pool revealed” , “ procedure” , “ misters showers fountains spa pools” , “ distribution systems” , “ cells starting” , “ condensers5”	**“ contaminated aerosols”** , “ cooling towers” , “ hot **water distribution systems”** , **“ humidifiers”** , “ misters” , “ showers” , “ fountains” , “ spa pools” , **“ evaporative condensers”**
Route of trans/on	“ tract” , “ micro aspiration” , “ inhalation”	“ inhalation” , “ micro aspiration”	“ aspiration” , “ inhalation” , “ million”	“ inhalation” , “ aspiration”	“ attempted” , “ million” , “ percentage” , “ inhaled” , “ modification”	“ inhaled”
Env. habitat	“ water samples” , “ water systems” , “ fixtures” , “ human”	“ water system”	“ hot water systems” , “ water samples” , “ hot and cold” , “ hot water sampling taps” , “ storage” , “ **plumbing systems** distributed municipal” , “ production” , “ municipal” , “ absorber” , “ system” , “ water networks”	“ water systems’ , “ water networks”	“ precipitation” , “ systems2” , “ water systems” , “ manmade engineered” , “ artificial freshwater” , “ plumbing” , “ manmade” , “ engineered water systems” , “ distribution systems”	“ natural and engineered water systems” , “ man-made engineered water systems”

	Interactive Effects of Copper Pipe, Stagnation, Corrosion Control,…(Martin et al., 2020)		Large community-acquired Legionnaires' disease outbreak caused by Legionella pneumophila serogroup 1(Faccini et al., 2020)		Legionella Diversity and Spatiotemporal Variation in the Occurrence of Opportunistic Pathogens within a Large Building Water System(Buse et al.,	
	System	**Manual extraction**	**System**	**Manual extraction**	**System**	**Manual extraction**
Incubatio	-	-	“ 2 and 10 days”	“ 2 and 10 days”	-	-
Species	“ L pneumophila”	“ L pneumophila”	“ legionella pneumophila”	“ legionella pneumophila”	“ gormanii” ,“ massiliensis” , “ feeleii” , “ longbeachae” , “ bozemanii” , “ nautarum”	“ gormanii” ,“ massiliensis ” , “ feeleii” , “ longbeachae”
Symptoms	-	-	-		-	
Cl. Manif/on	“ pipe material” , “ administration” , “ legionnaires disease ld”	“ Legionnaires disease”	“ pneumonia” , “ ld ranges” , “ legislation” , “ programmes” , “ legionnaires disease” , “ questionnaires”	“ Legionnaires ‘ disease” , “ pneumonia”	“ without passage” , “ legionellosis” , “ log10 ce” , “ processed” , “ ia” , “ pneumonia legionnaires disease ld”	“ legionellosis” , “ **Pontiac Fever”** , “ Legionnaires disease”
Source of exposure	“ reproduced” , “ material” , “ simulated distribution” , “ asks” , “ validation” , “ distribution” , “ plumbing” , “ summer” , “ water supply” , “ pipe detroit tap water” , “ biofilms” , “ working” , “ tap water” , “ distribution system” , “ antimicrobial properties” , “ system pipe” , “ mixing representing” , “ tap water euents” , “ water treatment raw”	“ plumbing” , “ tap water” , “ distribution system”	“ fountain site” , “ spa pools” , “ outdoor bioaerosols” , “ cooling tower” , “ water samples” , “ unit” , “ contaminated water” , “ bioaerosol” , “ precipitation” , “ devices” , “ multidisciplinary team” , “ cooling systems sprinklers” , “ aerosol production” , “ heavy rainfalls” , “ cooling towers evaporative condensers” , “ decorative fountains” , “ aerosol producing devices” , “ contaminated particles” , “ ascertain”	“ contaminated water systems” , “ cooling towers” , “ evaporative condensers” , “ spa pools” , “ cooling systems” , “ sprinklers” , “ fountains” , “ **municipal water**” , “ bioaerosols”	“ draw hot” , “ bws building water” , “ aerosolization into respirable “ agglutination” , “ bws supply” , “ tanks” , “ cooling tower” , “ contaminated water chillers” , “ structure clonal expansion” ,, “ pipe material” , “ cooling tower water” , “ cold potable water” , “ water supply” , “ fall seasons” , “ distribution system” , “ aerosols” , “ fountain” , “ validation” , “ plumbing for” , “ correlations between” , “ acanthamoeba spp”	“ contaminated aerosols” , “ contaminated water chillers” ” hot and cold water”
Route of trans/on	-	-	“ adjusted” , “ inhalation” , “ between”	“ inhalation”	“ inhalation”	“ inhalation”
Env. habitat	“ hot water systems” , “ pipe surface” , “ mature” , **“ building plumbing water systems**” , “ water supply”	“ **river water”** , “ water supply” , “ hot water systems” , “ pipes” , “ biofilms”	“ unit epidemiological” , “ pipe” , “ highly” , “ municipal water” , **“ fountains**”	“ municipal water” , “ water pipes”	“ serotyping latex agglutination” , **“ building water”** , **“ hot water**” , “ engineered water systems” , “ distribution systems” , “ water quality parameter” , “ engineered environments” , “ amoeba” , “ waters” , “ plumbing systems” , “ biofilm” , “ storage”	“ engineered water systems” , “ building water systems” , “ drinking water distribution systems” ” sea” , “ freshwater” , “ rainwater” , “ treated water” , “ biofilm”

	Legionellosis incidents associated with spa pools England (Dabrera et al., 2020)		Rapid Detection of Legionella pneumophila in Drinking Water, Based on Filter Immunoassay and Chronoamperometric Measurement (Ezenarro et al., 2020)	
	System	**Manual extraction**	**System**	**Manual extraction**
Incubation period	-	-	-	-
Species	-	-	“ L pneumophila”	“ L pneumophila”
Symptoms	-	-	“ **weakness”**	-
Cl. Manifestation	“ pf” , “ Pontiac fever” , “ legionnaires disease” , “ legionellosis” , “ “ pneumonia”	“ legionellosis” , “ Legionnaires disease” , “ Pontiac fever” , “ pneumonia”	“ Legionnaires disease ld” , “ biotica” , “ ranges” , “ pontiac fever”	“ Legionnaires disease” , Pontiac fever”
Source of exposure	“ users” , “ aerosols” , “ spa pools”	“ spa pools” , “ aerosols”	“ environments” , “ cells” , “ aerosols” , “ labourintensive timeconsuming” , “ **peristaltic pump” ,** “ showers cooling towers”	“ showers” , “ cooling towers**” , whirlpool spas” ,** “ aerosols”
Route of transmission	“ inhaled”	“ inhaled”	“ micro aspiration” , “ inhalation”	“ inhalation” , “ micro-aspiration”
Env. habitat	-	-	“ highly” , “ environmental water samples” , “ freshwater”	“ freshwater environments” , **man-made water systems**

References for 50 scientific publications related to Legionella

Amemura-Maekawa, J., Kura, F., Chida, K., Ohya, H., Kanatani, J.-I., Isobe, J., Tanaka, S., Nakajima, H., Hiratsuka, T., Yoshino, S., Sakata, M., Murai, M., Ohnishi, M., 2018. Legionella pneumophila and other Legionella species isolated from legionellosis patients in Japan between 2008 and 2016. Am Soc Microbiol 84. https://doi.org/10.1128/AEM.00721-18

Atlas, R.M., 1999. Legionella: From environmental habitats to disease pathology, detection and control. Environ. Microbiol. 1, 283–293. https://doi.org/10.1046/j.1462-2920.1999.00046.x

Beauté, J., 2017. Legionnaires' disease in Europe, 2011 to 2015. Eurosurveillance 22, 1–8. https://doi.org/10.2807/1560-7917.ES.2017.22.27.30566

Bédard, E., Paranjape, K., Lalancette, C., Villion, M., Quach, C., Laferrière, C., Faucher, S.P., Prévost, M., 2019. Legionella pneumophila levels and sequence-type distribution in hospital hot water samples from faucets to connecting pipes. Water Res. 156, 277–286. https://doi.org/10.1016/j.watres.2019.03.019

Benin, A.L., Benson, R.F., Arnold, K.E., Fiore, A.E., Cook, P.G., Williams, L.K., Fields, B., Besser, R.E., 2002. An outbreak of travel-associated legionnaires disease and pontiac fever: The need for enhanced surveillance of travel-associated legionellosis in the United States. J. Infect. Dis. 185, 237–243. https://doi.org/10.1086/338060

Bennett, E., Ashton, M., Calvert, N., Chaloner, J., Cheesbrough, J., Egan, J., Farrell, I., Hall, I., Harrison, T.G., Naik, F.C., Partridge, S., Syed, Q., Gent, R.N., 2014. Barrow-in-Furness: A large community legionellosis outbreak in the UK. Epidemiol. Infect. 142, 1763–1777. https://doi.org/10.1017/S0950268813002483

Brooks, T., Osicki, R.A., Susan Springthorpe, V., Abdus Sattar, S., Filion, L., Abrial, D., Riffard, S., 2004. DETECTION AND IDENTIFICATION OF Legionella SPECIES FROM GROUNDWATERS. Taylor Fr. 67, 1845–1859. https://doi.org/10.1080/15287390490492449

Burillo, A., Pedro-Botet, M.L., Bouza, E., 2017. Microbiology and Epidemiology of Legionnaire's Disease. Infect. Dis. Clin. North Am. 31, 7–27. https://doi.org/10.1016/j.idc.2016.10.002

Buse, H.Y., Morris, B.J., Gomez-Alvarez, V., Szabo, J.G., Hall, J.S., 2020. Legionella Diversity and Spatiotemporal Variation in the Occurrence of Opportunistic Pathogens within a Large Building Water System. Pathog. 2020, Vol. 9, Page 567 9, 567. https://doi.org/10.3390/PATHOGENS9070567

Caicedo, C., Rosenwinkel, K.H., Exner, M., Verstraete, W., Suchenwirth, R., Hartemann, P., Nogueira, R., 2019. Legionella occurrence in municipal and industrial wastewater treatment plants and risks of reclaimed wastewater reuse: Review. Water Res. 149, 21–34. https://doi.org/10.1016/j.watres.2018.10.080

Cebrián, F., Montero, J.C., Fernández, P.J., 2018. New approach to environmental investigation of an explosive legionnaireś disease outbreak in Spain: Early identification of potential risk sources by rapid Legionella spp immunosensing technique. BMC Infect. Dis. 18, 1–8. https://doi.org/10.1186/s12879-018-3605-8

Chahin, A., Opal, S.M., 2017. Severe Pneumonia Caused by Legionella pneumophila. Infect. Dis. Clin. North Am. 31, 111–121. https://doi.org/10.1016/j.idc.2016.10.009

Coscollá, M., Fenollar, J., Escribano, I., González-Candelas, F., 2010. Legionellosis outbreak associated with asphalt paving machine, Spain, 2009. Emerg. Infect. Dis. 16, 1381–1387. https://doi.org/10.3201/eid1609.100248

Cosentini, R., Tarsia, P., Blasi, F., Roma, E., Allegra, L., 2001. Community-acquired pneumonia: Role of atypical organisms. Monaldi Arch. Chest Dis. - Pulm. Ser. 56, 527–534.

Cunha, B.A., Burillo, A., Bouza, E., 2016. Legionnaires' disease. Lancet 387, 376–385. https://doi.org/10.1016/S0140-6736(15)60078-2

Cunha, C.B., Cunha, B.A., 2017. Legionnaire's Disease Since Philadelphia: Lessons Learned and Continued Progress. Infect. Dis. Clin. North Am. 31, 1–5. https://doi.org/10.1016/j.idc.2016.10.001

Currie, S.L., Beattie, T.K., 2015. Compost and Legionella longbeachae : an emerging infection? Perspect. Public Health 135, 309–315. https://doi.org/10.1177/1757913915611162

Dabrera, G., Naik, F., Phin, N., 2020. Legionellosis incidents associated with spa pools, England, 2002–2018. Public Health 185, 232–234. https://doi.org/10.1016/J.PUHE.2020.05.061

Darby, J., Buising, K., 2008. Could it be Legionella? Aust. Fam. Physician 37, 812–815.

Den Boer, J.W., Euser, S.M., Brandsema, P., Reijnen, L., Bruin, J.P., 2015. Results from the national Legionella outbreak detection program, the Netherlands, 2002–2012. Emerg. Infect. Dis. 21, 1167–1173. https://doi.org/10.3201/eid2107.141130

Den Boer, J.W., Verhoef, L., Bencini, M.A., Bruin, J.P., Jansen, R., Yzerman, E.P.F., 2007. Outbreak detection and secondary prevention of Legionnaires' disease: A national approach. Int. J. Hyg. Environ. Health 210, 1–7. https://doi.org/10.1016/j.ijheh.2006.07.002

Den Boer, J.W., Yzerman, E.P.F., Schellekens, J., Lettinga, K.D., Boshuizen, H.C., Van Steenbergen, J.E., Bosman, A., Van Hof, S. Den, Van Vliet, H.A., Peeters, M.F., Van Ketel, R.J., Speelman, P., Kool, J.L., Marina, A.E., Spaendonck, C. Van, 2002. A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. Emerg. Infect. Dis. 8, 37–43. https://doi.org/10.3201/eid0801.010176

Dilger, T., Melzl, H., Gessner, A., 2018. Legionella contamination in warm water systems: A species-level survey. Int. J. Hyg. Environ. Health 221, 199–210. https://doi.org/10.1016/j.ijheh.2017.10.011

Essig, A., von Baum, H., Gonser, T., Haerter, G., Lück, C., 2016. Microbiological diagnosis and molecular typing of Legionella strains during an outbreak of legionellosis in Southern Germany. Int. J. Med. Microbiol. 306, 109–114. https://doi.org/10.1016/j.ijmm.2016.01.001

Ezenarro, J.J., Párraga-Niño, N., Sabrià, M., Del Campo, F.J., Muñoz-Pascual, F.X., Mas, J., Uria, N., 2020. Rapid Detection of Legionella pneumophila in Drinking Water, Based on Filter Immunoassay and Chronoamperometric Measurement. Biosens. 2020, Vol. 10, Page 102 10, 102. https://doi.org/10.3390/BIOS10090102

Faccini, M., Russo, A.G., Bonini, M., Tunesi, S., Murtas, R., Sandrini, M., Senatore, S., Lamberti, A., Ciconali, G., Cammarata, S., Barrese, E., Ceriotti, V., Vitaliti, S., Foti, M., Gentili, G., Graziano, E., Panciroli, E., Bosio, M., Gramegna, M., Cereda, D., Perno, C.F., Mazzola, E., Campisi, D., Aulicino, G., Castaldi, S., Girolamo, A., Caporali, M.G., Scaturro, M., Rota, M.C., Ricci, M.L., 2020. Large community-acquired Legionnaires' disease outbreak caused by Legionella pneumophila serogroup 1, Italy, July to August 2018. Eurosurveillance 25, 1900523. https://doi.org/10.2807/1560-7917.ES.2020.25.20.1900523/CITE/PLAINTEXT

Fenstersheib, M.D., Miller, M., Diggins, C., Liska, S., Detwiler, L., Werner, S.B., Lindquist, D., Thacker, W.L., Benson, R.F., 1990. Outbreak of Pontiac fever due to Legionella anisa. Lancet 336, 35–37. https://doi.org/10.1016/0140-6736(90)91532-F

Francois Watkins, L.K., Toews, K.E., Harris, A.M., Davidson, S., Ayers-Millsap, S., Lucas, C.E., Hubbard, B.C., Kozak-Muiznieks, N.A., Khan, E., Kutty, P.K., 2017. Lessons From an Outbreak of Legionnaires' Disease on a Hematology-Oncology Unit. Infect. Control Hosp. Epidemiol. 38, 306–313. https://doi.org/10.1017/ice.2016.281

Fujii, J., Yoshida, S.I., 1998. Legionella infection and control in occupational and environmental health. Rev. Environ. Health 13, 179–203. https://doi.org/10.1515/REVEH.1998.13.4.179

Hamilton, K.A., Prussin, A.J., Ahmed, W., Haas, C.N., 2018. Outbreaks of Legionnaires' Disease and Pontiac Fever 2006–2017. Curr. Environ. Heal. reports 5, 263–271. https://doi.org/10.1007/s40572-018-0201-4

Harris, L.F., Battersby, E.J., 1990. Cutaneous Legionella longbeachae Infection in Immunosuppressed Woman, United Kingdom. Ala. Med. 59, 8–10. https://doi.org/10.32388/sjt5i0

Herwaldt, L.A., Marra, A.R., 2018. Legionella: A reemerging pathogen. Curr. Opin. Infect. Dis. 31, 325–333. https://doi.org/10.1097/QCO.0000000000000468

Lanternier, F., Ader, F., Pilmis, B., Catherinot, E., Jarraud, S., Lortholary, O., 2017. Legionnaire's Disease in Compromised Hosts. Infect. Dis. Clin. North Am. 31, 123–135. https://doi.org/10.1016/j.idc.2016.10.014

Lee, S., & Lee, J., 2013. Outbreak Investigations and Identification of Legionella in Contaminated Water. Springer.

Lettinga, K.D., Verbon, A., Weverling, G.J., Schellekens, J.F.P., Den Boer, J.W., Yzerman, E.P.F., Prins, J., Boersma, W.G., Van Ketel, R.J., Prins, J.M., Speelman, P., 2002. Legionnaires' disease at a Dutch flower show: Prognostic factors and impact of therapy. Emerg. Infect. Dis. 8, 1448–1454. https://doi.org/10.3201/eid0812.020035

Leung, Y.H., Lam, C.K., Cheung, Y.Y., Chan, C.W., Chuang, S.K., 2020. Epidemiology of Legionnaires' Disease, Hong Kong, China, 2005−2015. Emerg. Infect. Dis. 26, 1695. https://doi.org/10.3201/EID2608.191244

Lucas, K.D., Wheeler, C., McLendon, P., Leistikow, B.N., Mohle-Boetani, J.C., 2018. Outbreak of Legionnaires' disease associated with cooling towers at a California state prison, 2015. Epidemiol. Infect. 146, 297–302. https://doi.org/10.1017/S0950268818000110

Martin, R.L., Strom, O.R., Pruden, A., Edwards, M.A., 2020. Interactive Effects of Copper Pipe, Stagnation, Corrosion Control, and Disinfectant Residual Influenced Reduction of Legionella pneumophila during Simulations of the Flint Water Crisis. Pathog. 2020, Vol. 9, Page 730 9, 730. https://doi.org/10.3390/PATHOGENS9090730

Moran-Gilad, J., Lazarovitch, T., Mentasti, M., Harrison, T., Weinberger, M., Mordish, Y., Mor, Z., Stocki, T., Anis, E., Sadik, C., Amitai, Z., Grotto, I., 2012. Huhumidifier-associated paediatric legionnaires' disease, Israel, February 2012. Eurosurveillance 17, 1–4. https://doi.org/10.2807/ese.17.41.20293-en

Mouchtouri, V.A., Rudge, J.W., 2015. Legionnaires' Disease in Hotels and Passenger Ships: A Systematic Review of Evidence, Sources, and Contributing Factors. J. Travel Med. 22, 325–337. https://doi.org/10.1111/jtm.12225

Parr, A., Whitney, E.A., Berkelman, R.L., 2015. Legionellosis on the rise: A review of guidelines for prevention in the United States. J. Public Heal. Manag. Pract. 21, E17–E26. https://doi.org/10.1097/PHH.0000000000000123

Phin, N., Parry-Ford, F., Harrison, T., Stagg, H.R., Zhang, N., Kumar, K., Lortholary, O., Zumla, A., Abubakar, I., 2014. Epidemiology and clinical management of Legionnaires' disease. Lancet Infect. Dis. 14, 1011–1021. https://doi.org/10.1016/S1473-3099(14)70713-3

Reuter, S., Harrison, T.G., Köser, C.U., Ellington, M.J., Smith, G.P., Parkhill, J., Peacock, S.J., Bentley, S.D., Török, M.E., 2013. A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. BMJ Open 3, 1–6. https://doi.org/10.1136/bmjopen-2012-002175

Saad, M., Chinerman, D., Tabrizian, M., Faucher, S.P., 2020. Identification of two aptamers binding to Legionella pneumophila with high affinity and specificity. Sci. Reports 2020 101 10, 1–10. https://doi.org/10.1038/s41598-020-65973-3

Scaturro, M., Fontana, S., Crippa, S., Caporali, M.G., Seyler, T., Veschetti, E., Villa, G., Rota, M.C., Ricci, M.L., 2014. An unusually long-lasting outbreak of community-acquired Legionnaires' disease, 2005-2008, Italy. Epidemiol. Infect. 128, 2416–2425. https://doi.org/10.1017/S0950268814003094

Schnirman, R., Nur, N., Bonitati, A., Carino, G., 2017. A case of legionella pneumonia caused by home use of continuous positive airway pressure. SAGE Open Med. Case Reports 5, 2050313X1774498. https://doi.org/10.1177/2050313x17744981

Schönning, C., Jernberg, C., Klingenberg, D., Andersson, S., Pääjärvi, A., Alm, E., Tano, E., Lytsy, B., 2017. Legionellosis acquired through a dental unit: a case study. J. Hosp. Infect. 96, 89–92. https://doi.org/10.1016/j.jhin.2017.01.009

Schwake, D.O., Alum, A., Abbaszadegan, M., 2015. Automobile windshield washer fluid: A potential source of transmission for Legionella. Sci. Total Environ. 526, 271–277. https://doi.org/10.1016/j.scitotenv.2015.03.122

Smith, A.F., Huss, A., Dorevitch, S., Heijnen, L., Arntzen, V.H., Davies, M., van Beest Holle, M.R.D.R., Fujita, Y., Verschoor, A.M., Raterman, B., Oesterholt, F., Heederik, D., Medema, G., 2019. Multiple sources of the outbreak of legionnaires' disease in Genesee County, Michigan, in 2014 and 2015. Environ. Health Perspect. 127. https://doi.org/10.1289/EHP5663

Springston, J.P., Yocavitch, L., 2017. Existence and control of Legionella bacteria in building water systems: A review. J. Occup. Environ. Hyg. 14, 124–134. https://doi.org/10.1080/15459624.2016.1229481

Thornley, C.N., Harte, D.J., Weir, R.P., Allen, L.J., Knightbridge, K.J., Wood, P.R.T., 2017. Legionella longbeachae detected in an industrial cooling tower linked to a legionellosis outbreak, New Zealand, 2015; possible waterborne transmission? Epidemiol. Infect. 145, 2382–2389. https://doi.org/10.1017/S0950268817001170

Totaro, M., Costa, A.L., Frendo, L., Profeti, S., Casini, B., Gallo, A., Privitera, G., Baggiani, A., 2020. Evaluation of Legionella spp. Colonization in Residential Buildings Having Solar Thermal System for Hot Water Production. Int. J. Environ. Res. Public Heal. 2020, Vol. 17, Page 7050 17, 7050. https://doi.org/10.3390/IJERPH17197050

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

Ward, M., Boland, M., Nicolay, N., Murphy, H., McElhiney, J., Collins, C., Lynch, M., McCarthy, M., O'Donnell, J., 2010. A cluster of legionnaires' disease and associated pontiac fever morbidity in office workers, Dublin, June-July 2008. J. Environ. Public Health 2010. https://doi.org/10.1155/2010/463926

Whiley, H., Bentham, R., 2011. Legionella longbeachae and legionellosis. Emerg. Infect. Dis. 17, 579–583. https://doi.org/10.3201/eid1704.100446

Wüthrich, D., Gautsch, S., Spieler-Denz, R., Dubuis, O., Gaia, V., Moran-Gilad, J., Hinic, V., Seth-Smith, H.M.B., Nickel, C.H., Tschudin-Sutter, S., Bassetti, S., Haenggi, M., Brodmann, P., Fuchs, S., Egli, A., 2019. Air-conditioner cooling towers as complex reservoirs and continuous source of Legionella pneumophila infection evidenced by a genomic analysis study in 2017, Switzerland. Eurosurveillance. https://doi.org/10.2807/1560-7917.ES.2019.24.4.1800192

Xiong, L., Yan, H., Shi, L., Mo, Z., 2016. Antibiotic susceptibility of Legionella strains isolated from public water sources in Macau and Guangzhou. J. Water Health 14, 1041–1046. https://doi.org/10.2166/wh.2016.056

Zanella, M.C., Yerly, S., CherkaouiI, A., Renzi, G., Mamin, A., Lourenço Cordes, L., Delaporte, E., Baranczuk-Turska, Z., Keiser, O., Schrenzel, J., Harbarth, S., Gaia, V., Kaiser, L., 2018. A community outbreak of Legionnaires' disease in Geneva, Switzerland, June to September 2017. Swiss Med. Wkly. 148, w14687. https://doi.org/10.4414/smw.2018.14687

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. Informatics 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 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. Inform. 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. Inform. Decis. Mak. 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), 2019. A Deep Neural Network Model for the Task of Named Entity Recognition. Int. J. Mach. Learn. Comput. 9.

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. Recognit. 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. Public Heal. 2018, Vol. 15, Page 598 15, 598. https://doi.org/10.3390/IJERPH15040598

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

Uzuner, Ö., Solti, I., Cadag, E., 2010. Extracting medication information from clinical text. J. Am. Med. Informatics 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. Amsterdam: Centrum voor Wiskunde en Informatica.

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.

Chapter 3

Extracting information about drinking water contamination events from the scientific literature: performance evaluation of Large Language Models

Submitted in the Journal of Environmental Informatics Letters. Authors: Paraskevopoulos, S., Ribalta, M., Ravishankara, C., Smeets, P., & Medema, G.

Chapter 3- Extracting information about drinking water contamination events from the scientific literature: performance evaluation of Large Language Models

Abstract

During pathogen contamination events of drinking water, water utilities want access to the scientific literature, providing them with information and insights derived from actual (past) contamination events so they can rapidly assess the situation and respond accordingly. The need for rapid access to information suggests that manual literature reviews might not be sufficient and necessitate the use of Artificial Intelligence and specifically Large Language Models to gather such information from text documents.

This study systematically assessed the performance of multiple open-source Large Language Models, including Llama 2, Mistral, and Gemma (and their variations) in a question-answering task related to pathogen contamination events of drinking water. The evaluation metrics included Precision, Recall, F1 score, Automated Accuracy, and Empty Score. The model with the highest performance on a set of 23 questions (prompts) using 188 scientific publications was then manually evaluated by a human.

The results showed that all models performed reasonably well with an average F1 score ranging from 81% to 87%. After considering all the evaluation metrics, the Llama 2 model was the most reliable model with an average Automated Accuracy of 86%. However, the hallucination effect of Llama 2 was evident, unlike the Gemma model, which although it had a lower Automated Accuracy score, it was less prone to hallucination. The Human Evaluation showed that the Llama 2 model delivered correct answers when the questions were clear and straightforward. However, when the question required further interpretation, the model often struggled. The use of LLMs in automated information extraction tasks shows great potential for time-critical applications, such as processing large volumes of (historical) data in real-time thereby reducing the time required for manual literature reviews in case of emergencies. Further development of hybrid models that can combine strengths from multiple LLMs could lead to robust, high-accuracy information extraction tools for emergency management and decision-support in the drinking water sector.

Keywords: Contamination events; Large Language Models; Information Extraction; Question-answering; Emergencies

3.1 Introduction

During pathogen contamination events of drinking water (DW), water utilities want to ensure the safety of drinking water and rapid restoration of the water supply system since human exposure to pathogens poses serious health risks (Hrudey and Hrudey, 2004). Accurate and timely information is crucial during such crises to rapidly assess and respond. As serious contamination events of DW are rare, water utilities have limited experience with how such an event unfolds, what the health impact (and impact on customer trust) can be, what effective mitigation strategies are etc. The initial information at the onset of an event is generally limited, while the need to respond is urgent. There are response protocols in place, but there is little access to the scientific literature that would allow to gather information and insights derived from actual (past) pathogen contamination events. These insights often cover the cause and source of contamination, the associated pathogens and their typical concentrations, the number of people exposed, sick, or deceased, as well as any monitoring and prevention measures implemented by the responsible authorities, and how effective they have been.

The need for a timely response, the diverse nature of pathogens, and the complexity of how pathogen contamination events are reported in the scientific literature suggest that traditional methods, such as manually performing literature reviews and meta-analyses, might not be sufficient (Paraskevopoulos et al., 2022) when decision-makers need instant access to information. Artificial Intelligence (AI), which provides rapid data processing and information extraction (IE) capabilities, can revolutionize responses to emergencies related to pathogen contamination events in drinking water supply.

Over the past two decades, AI has emerged as a powerful tool for automating information extraction. Biomedical sciences, in particular, have leveraged AI-driven Information Extraction techniques, especially for extracting data from Electronic Health Records and Randomized Control Trials. Common approaches, include Machine Learning (ML), Deep Learning (DL) and Natural Language Processing (NLP) as illustrated by studies like Cohen and Hersh (2005), Meystre et al. (2008), Kiritchenko et al. (2008), Patrick and Li (2010). Paraskevopoulos et al. (2022) tested the extraction of water pathogen information from the scientific literature using DL, NLP and rule-based techniques and developed a model that could efficiently extract such information with sufficient accuracy.

In the scientific literature of outbreaks and pathogen-related emergencies in drinking water supply, the complexity and unstructured information in text presents a challenge for IE and QA tasks that traditional techniques (in ML and DL) cannot properly handle.

A significant architecture within DL for processing sequential data, like text, is the Transformer model, which uses self-attention mechanisms to dynamically weigh the importance of different inputs (Vaswani et al., 2017). This architecture is foundational in modern language processing tasks, including question-answering (QA) tasks (Rajpurkar et al., 2016). For example, DL techniques use models that can digest large volumes of text (structured or unstructured) and provide answers to questions (queries) by understanding context and relevance. Historically, Transformer architectures, such as BERT (Bidirectional Encoder Representations from Transformers), have been influential in the field of NLP. Developed by researchers at Google AI and introduced in 2018 through the paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (Devlin, 2018), BERT marked a significant advancement in how machines understand human language.

Large Language Models (LLMs) are a type of transformer, based on a Decoder architecture, created to understand and generate human language from raw text. They are considered "large" because they can process vast amounts of text data and have architectures that include billions of parameters. LLMs can handle various language tasks, such as translation, summarization, question answering, and text generation, often with minimal task-specific adjustments. Recently, LLMs have become crucial for IE and QA tasks excelling in interpreting complex medical texts and retrieving accurate information from diverse sources (Kartchner et al., 2023; Chen et al., 2023).

The shift from BERT-like models to contemporary LLMs has been characterized by increasing the model size (more layers, more parameters), expanding training data (more diverse, larger datasets), and enhancing versatility (capable of more varied tasks). This progression reflects a broader trend in AI towards building more powerful, efficient, and adaptable systems.

Given the impressive performance of LLMs in performing IE and QA tasks across general and biomedical sciences corpora, there is a promising potential for their application in the environmental microbiology scientific field. Although LLMs have been successfully used in the water sector to improve environmental predictions (Li et al., 2024b), or simplify complex analytical processes (Liang et al., 2024), their performance on QA tasks in environmental microbiology has not been tested.

This Chapter aims to fill that gap by systematically assessing the performance of multiple open-source LLMs, including Llama 2 (7b), Mistral (7b), and Gemma (2b and 7b) on a QA task related to pathogen contamination events of drinking water. The goal of this Chapter is to identify the LLM with the highest performance using multiple evaluation metrics and to manually evaluate the top-performing model's answers on a set of 23 pathogen-related questions (prompts), to determine what information can be extracted with sufficient accuracy.

3.2 Methodology

To evaluate the performance of different LLMs on a QA task, we first gathered scientific publications related to outbreaks and pathogen-related contamination events of DW. After collecting the data, we processed and formatted the publications to ensure consistency. The data was then annotated for the QA task applied to each LLM. Figure 3.1 provides the flowchart of this QA task. All the evaluation results along with the questions and raw data can be found in our Github repository (Ribalta, 2024).

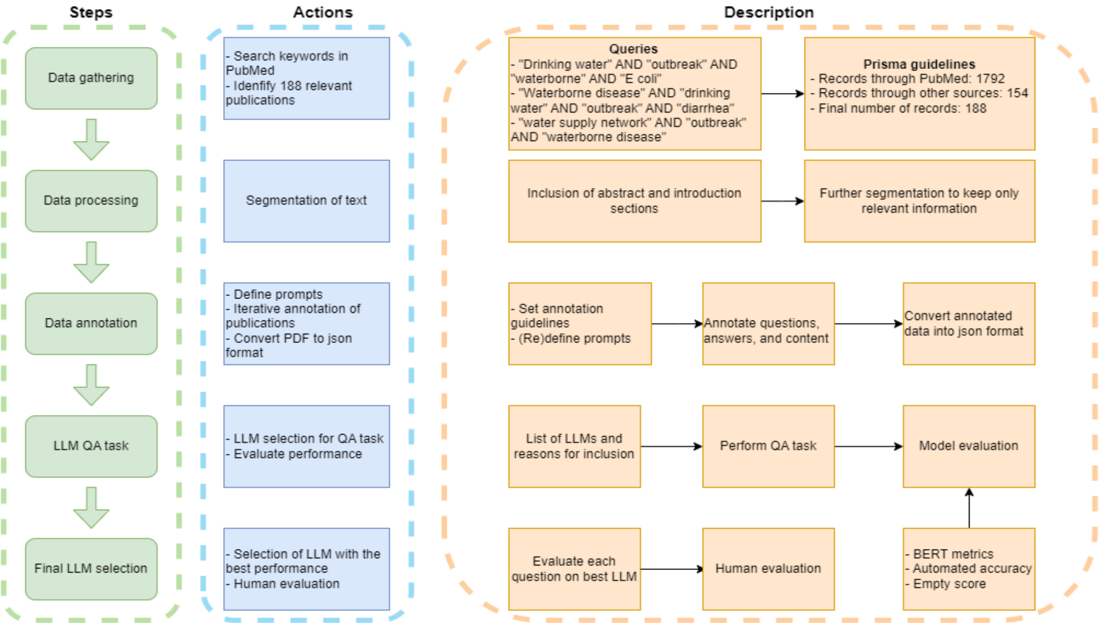


Figure 3. 1 The flowchart of the QA task including the steps, the actions, and the description.

3.2.1 Data gathering

The scientific publications were collected through a systematic review of the literature using the search engine PubMed. The full list of the search query can be found in Figure S3.1 of the supplementary material. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Liberati et al., 2009) was adopted. The selection of publications was based on their relevance to pathogen contamination events occurring in drinking water and their ability to answer the maximum number of questions in the QA task. The publications were evaluated based on specific criteria such as topic comprehensiveness of the information, and clarity on the desired information. Ultimately, a total of 188 publications were selected addressing outbreaks, infrastructure failures, and accidents that eventually led to pathogen contamination events in drinking water, as well as their associated response measures. The entire flowchart of the PRISMA guidelines can be found in Figure S3.2 of the supplementary material.

3.2.2 Data processing

A manual check was performed on a set of publications to ensure consistency in section headings across all publications. Based on these checks, we assumed consistency and used rule-based keyword matching techniques to automatically remove sections irrelevant to the QA task, including the methodology, discussion, conclusion, acknowledgements, and references sections. This approach ensured that the filtering process was reliable, despite any minor inconsistencies. The relevant information, such as cause and source of contamination, pathogen concentrations, number of people exposed, sick, or deceased, and mitigation measures were typically found in the abstract, introduction and results sections.

The extracted information from each publication was then converted into .txt format to be processed as raw data. Subsequently, the processed publications were further segmented to keep only the sentences and content relevant to the desired information. This way the model does not have to parse unnecessary text in the abstract and introduction sections. Figure S3 in the supplementary material shows an example of this process, highlighting sentences with relevant information within the introduction section of a publication.

3.2.3 Data annotation

In this step, an annotated dataset containing each publication, the associated questions and their corresponding answers, was created. This annotated dataset served as the basis for evaluating and comparing the performance of different LLMs in the QA task.

To ensure consistency, clear annotation guidelines were set including how to identify the correct information in the text, handle ambiguous cases, and deal with overlapping information. For example, if a question cannot be answered directly from the text, it is up to the annotator (an expert) to find relevant text that indirectly answers the question or to decide whether there is no information at all. Additionally, in cases where the text is ambiguous and the annotator is not sure of what is factually correct versus what is explicitly stated in the text, the annotator should prioritize what humans would perceive as correct, rather than trying to anticipate what the model might interpret as correct.

We then developed 23 questions related to pathogen contamination events in drinking water. These questions were designed to test different aspects of the LLMs' comprehension and retrieval capabilities. Some of them were simple questions with straightforward answers such as: "What was the date of the event?", and "What was the location of the event?". Other questions were more ambiguous in the sense that sometimes it would be difficult even for a human to find an answer in the text segment. Some of these questions include: "What was the cause/source of contamination?", and "What were the mitigation measures?". Table 3.1 provides the entire list of 23 questions along with a description of each one of them and 4 main topics of questions. As can be seen, certain questions instruct the model not to infer answers. This is because, due to the model’s tendency to hallucinate, some questions had to be redefined after we observed that the answers were vague, or the model provided irrelevant responses.

Table 3.1. The list of questions (prompts) to be answered by the LLMs.

Topic	Questions
Event details	1. When did the contamination event occur?
	2. Where did the contamination event take place?
	3. Can you provide a summary of the contamination event?
	4. In the text, what triggered the contamination event? Do not infer an answer.
	5. What was identified as the origin of the contamination?
	6. How was the contamination event first discovered?
	7. Over what period did the contamination event span?
Impact	8. How many individuals were exposed to the contaminant? Don't write the illness count.
	9. What was the number of people who fell ill due to the event?
	10. What is the ratio of individuals who became ill to those who were exposed?
	11. How many fatalities were associated with the contamination event?
	12. What symptoms were associated with the contamination?
	13. What are the age demographics of the individuals mentioned within the text?
Investigation	14. What pathogens were found in the collected samples?
	15. What were the initial actions taken to investigate or inspect the event?
	16. Can you describe the detailed investigation or inspection?
	17. What is the case definition used?
	18. Which risk factors were identified during the investigation?
	19. What types of water analyses methods were performed during the monitoring phase?
	20. Name the contaminants detected using water sampling analysis. Do not infer positive detections.
Mitigation	21. What immediate actions were taken to mitigate the effects of the contamination?
	22. What steps were taken to restore the system after some time passed? Only say those mentioned in the text.
	23. What measures have been implemented to prevent future contamination events?

The manual annotation was then performed by reading through the segmented text (Data processing step) from each publication where for each question, the annotator identified and highlighted the relevant answer.

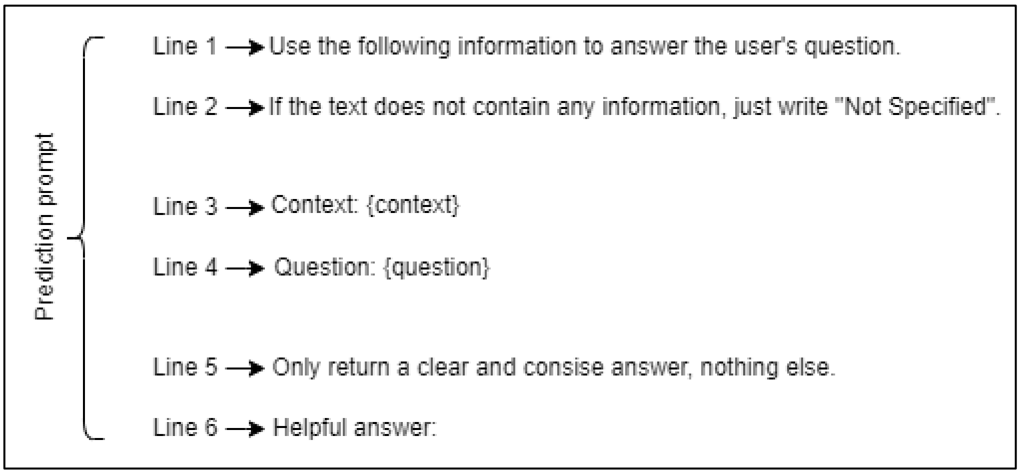
Finally, all the annotated data were gathered into a single JSON file, where each entry contained the publication ID and title, the relevant text segment (usually one or two sentences), the question, and the annotated answer. The JSON file containing this structured dataset was then used in the QA task, providing a reliable basis for evaluating the LLMs' performance.

3.2.3.1 (Re)defining prompts

As can be seen in Table 3.1, certain questions instruct the model not to infer answers. This is because, due to the model's tendency to hallucinate, some questions had to be redefined after we observed that the answers were vague, or the model provided irrelevant responses. Hallucination refers to the generation of text that is factually incorrect or nonsensical but presented in a manner that appears plausible and coherent.

The hallucination effect was also tackled in the creation and optimization of the prompt for each LLM. Figure 3.2 shows the final prompt that was used to feed each LLM model before the QA task. It is divided into different parts, one per line:

- First line: Instruct the model to use the provided information to answer the given question.
- Second line: Instruct the model on how to proceed if there is no relevant response in the context.
- Third line: Specify the text of the file that must be extracted.
- Fourth line: State the specific question out of the 23 needed.
- Fifth line: Remind the model not to generate any invented answers.
- Sixth line: Guide the model to start answering directly, ensuring it does not include an introductory part in the response and answers without added syntactic sugar.



Line 1 → Use the following information to answer the user's question.
Line 2 → If the text does not contain any information, just write "Not Specified".
Line 3 → Context: {context}
Line 4 → Question: {question}
Line 5 → Only return a clear and consise answer, nothing else.
Line 6 → Helpful answer:

Figure 3.2. The final prompt before each QA task for all LLMs.

3.2.4 LLM QA task

3.2.4.1 LLM selection

Although existing benchmarks of the most prominent LLMs provide insights into their performance under generic conditions, their relevance and effectiveness specifically within the environmental microbiology field remain to be determined. To select the most suitable LLMs for the QA task, various pre-trained models were evaluated. It was essential for the models to meet computational requirements that would allow for local storage within an organization. This criterion was established to avoid reliance on cloud services, which typically operate on a pay-per-subscription or pay-per-use basis.

The chosen LLMs are among the most well-known in the state of the art, with similar performance benchmarks. Different sizes were chosen for each type of LLM to compare performance under different computational costs. The selected LLMs are:

- LLaMa 2 7b[12] (Touvron et al., 2023): LLaMa is a series of models developed by Meta. It is known for its versatility and robust performance in multiple NLP benchmarks, and it is suitable for a variety of NLP tasks such as text generation, translation, summarization, and question-answering. The “7b” in LLaMa 2 7b denotes that the model has 7 billion parameters.
- Mistral 7b[13] (Jiang et al., 2023): Developed by the Mistral AI team, it is claimed to outperform the LLaMa 2 13b model across all benchmarks. Released under the Apache 2.0 license, it provides a strong performance with an emphasis on efficiency and open-source accessibility.
- Gemma 7b[14] (Team et al., 2023): Part of the Gemma family models developed by the Google Gemini team. This model is also claimed to be superior to LLaMa 2 13b in all benchmarks. It offers advanced capabilities in NLP tasks, benefiting from Google’s extensive research and development in AI.

---

[12] https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGUF/tree/main
[13] https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF/tree/main
[14] https://huggingface.co/mlabonne/gemma-7b-it-GGUF/tree/main

- Gemma 2b[15] (Team et al., 2023): Another model from the Gemma family, also developed by the Google Gemini team. As one of the first models to compete at the state of the art with only 2 billion parameters, it was included to understand the potential impact of smaller models on a possible growing trend.

LLMs can be slightly modified to decrease their computational cost with a small loss in performance. For this, the computational cost of the models was checked and reduced when Quantization was applied. Quantization involves converting the floating-point numbers (typically 32-bit or 16-bit) used in model weights and activations to lower bit-width representations, such as 8-bit integers. The main goal of quantization is to make the model more efficient in terms of memory usage, computational power, and energy consumption, while maintaining an acceptable level of accuracy. The variations of the models used in the study were limited to those available: Gemma 7b Q8 (8-bit integers Quantization), Mistral 7b Q8, and Llama 2 7b Q8 and Q4 (4-bit integers Quantization).

3.2.4.2 Evaluation metrics

The performance of the various LLMs was evaluated considering two important factors: that the answer is contextually correct, and that the model does not hallucinate. A correct context means that even though the model's answer is written differently than in the annotated text, the meaning is the same. Hallucination has already been explained in Chapter 3.2.3.1. The evaluation metrics used to evaluate these two factors are:

- BERT score (Precision, Recall, F1): A metric that uses the pre-trained contextual embeddings from BERT to evaluate the quality of text generation by comparing it with a reference (annotated) text using cosine similarity. More information on how the BERT score works can be found in Zhang et al. (2019).
    - Precision: Measures how many tokens (or embeddings) from the predicted text are close matches (semantically similar) to tokens in the reference text (quality of prediction).
    - Recall: Measures how many tokens from the reference text are captured by the predicted text (reference coverage).
    - F1 score: The harmonic mean of precision and recall. It balances the two metrics, providing a single number that considers both how much of the predicted text matches the reference and how much of the reference is covered by the prediction.
- Empty score: The 23 questions cannot be answered across all 188 scientific publications simply because the relevant information is not present in all of them. Models should recognize this and indicate that the document does not contain any answer. This metric measures how many times (percentage) the model accurately identifies such instances. Only this and the Human Evaluation metric (see below) assess the model's ability to correctly predict empty responses when there is no information available. The remaining metrics (BERT Score, Automated Accuracy) focus solely on the extraction of information.
- Automated Accuracy: This metric automatically measures how often the model being evaluated correctly answers a question. The process uses a voting system where the prediction accuracy of the model being evaluated (evaluatee), is assessed by the remaining models (evaluators). Each LLM evaluator is asked with a prompt to vote with a simple "Yes" or "No" on whether the predicted answer contains the same contextual information as the annotated answer, and the majority vote determines the final decision. Below is a chunk of the code that generates the prompt for the voting system. The full implementation of the voting system can be found in the GitHub repository[16]:

---

[15] https://huggingface.co/google/gemma-2b-it/tree/main

[16] https://github.com/Applied-Artificial-Intelligence-Eurecat/PathoTHREAT-data-retrieval/blob/main/src/auto_accuracy/evaluate.py

prompt="""

I will provide two text strings below. Do these two strings have the same contextual information?

After your analysis, respond with 'yes' if they share the same context, or 'no' if they do not.

String 1: {candidate}

String 2: {reference}

Answer only yes or no. Further wording is useless.

The answer is:

"""

- Human Evaluation: The real accuracy of the best-performing model is then calculated manually. This metric is the manual measurement of the percentage of times the top-performance model correctly answers a question. For every question, the prediction is reviewed manually by a human, one by one, for each of the 188 scientific publications. Due to the time-intensive nature of this manual activity, it is performed only for the best-performing model, which is selected based on the other three metrics: BERT Score, Empty Score, and Automated Accuracy.

3.3 Results

3.3.1 LLM evaluation

Table 3.2 presents the performance of each LLM on the QA task for a specific publication. Although only five questions and answers are shown here, the responses to all questions for all 188 scientific publications for each model are available in the GitHub repository (Ribalta, 2024). As can be seen, there is high variability in how each LLM responds to the questions. While these results are limited to a single publication and conclusions cannot be drawn, it is interesting to observe the differences in the reasoning capabilities of each model.

Table 3.2 Examples of responses from each LLM to selected questions in the QA task. The selected publication was from Martin et al. (2006).

LLM	Q9. What was the number of people who fell ill due to the event?	Q10. What is the ratio of individuals who became ill to those who were exposed (attack rate)?	Q22. What steps were taken to restore the system after some time passed?
Gemma 2b	Not Specified	The attack rate of AGI was 2.3 times higher in those who consumed water than others (AR = 27.3%).	Not Specified
Gemma 7b Q8	Not Specified	The attack rate of AGI was highest among those who consumed water (AR = 18%). The attack rate was also high among those who consumed communal water.	Not Specified
Mistral 7b Q8	The number of people who fell ill due to the event was 6-7000. However, only a subset of these cases were laboratory confirmed as Campylobacter infection or AGI. The exact number of these cases is not mentioned in the provided context.	The attack rate of AGI is the ratio of individuals who became ill to those who were exposed. In this study, the attack rate of AGI was compared by the amount of water consumed. The chi squared value for trend was used.	Not Specified
Llama 2 7b Q4	The number of people who fell ill due to the event was 6-7000.	The attack rate for individuals who consumed water was 27.3% (95% CI: 23.1-31.5%).	The water supply was intensified with chlorination of the water supplies of the municipality; and holding a press conference to inform the population. No advice on boiling water was issued.
Llama 2 7b Q8	The number of people who fell ill due to the event was 6-7000.	The attack rate for individuals who became ill was 27.3% (95% CI: 23.1-31.5%).	The epidemiological investigations included intensification of water sampling; chlorination of the water supplies of the municipality; holding a press conference to inform the population; and no advice on boiling water was issued.

Continuation of Table 3.2 Examples of responses from each LLM to selected questions in the QA task. The selected publication was from Martin et al. (2006).

LLM	Q2. Where did the contamination event take place?	Q6. How was the contamination event first discovered?
Gemma 2b	The contamination event took place in Soderhamn municipality.	The text does not specify how the contamination event was first discovered.
Gemma 7b Q8	Not Specified	The contamination event was first discovered through an increased number of domestically verified cases of campylobacter infection as well as a large outbreak of acute gastroenteritis infection (AGI) that was concurrently occurring within this time frame within this particular community of interest
Mistral 7b Q8	The contamination event took place at a water supply source in the Soderhamn municipality during December 2002 and January 2003.	The contamination event was first discovered through increased reports of domestic, laboratory confirmed cases of Campylobacter infection in the municipality of Soderhamn during December 2002 and January 2003. Concurrently, during preliminary investigations a large outbreak of acute gastroenteritis infection (AGI) was detected. The contamination event was first discovered through routine water sampling by Gavle Hospital and SMI laboratories.
Llama 2 7b Q4	The contamination event took place in the water supply of the municipality of Soderhamn.	The contamination event was first discovered through routine water sampling by Gavle Hospital and SMI laboratories.
Llama 2 7b Q8	Soderhamn municipality.	The contamination event was first discovered through a rapid assessment based on 750 telephone interviews indicating that 24% of the population self-reported AGI, potentially affecting some 6-7000 members of the population.

The results of evaluating the different models can be seen in Figure 3.3 where the models have an average F1 score ranging from 81% to 87%. Various observations can be made from these results, such as the small differences between the models in the Precision, Recall, and F1 metrics. When comparing these metrics with the Automated Accuracy, we notice discrepancies that are not reflected in the metrics provided by the BERT Score metric. Specifically, the Gemma models perform significantly worse than the Llama or Mistral models. Another notable observation is that while the Gemma models effectively predict instances where the answer is not included in the documents (76-80%), this ability is not observed in the Mistral and Llama models (12-25%). It is visually evident that models with better Automated Accuracy perform worse in predicting empty records. This happens because both Mistral and Llama models tend to hallucinate more frequently compared to the Gemma model when faced with unclear prompts. Therefore, when a question cannot be answered (because there is no apparent answer in the text), these models will provide a plausible yet incorrect answer.

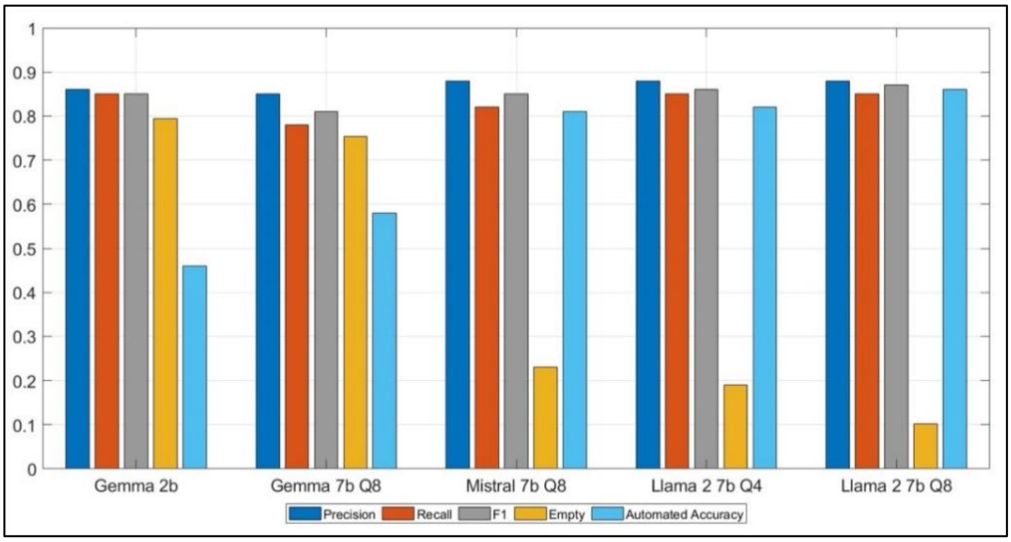


Figure 3.3. The performance of each LLM in the QA task.

3.3.1.1 Best LLM selection

To choose the best model, all metrics have been considered. Despite Llama 2 quantized to 8 integers having the worst performance in the Empty Score metric, it was selected for manual evaluation because it achieves the best results in the remaining evaluation metrics, including an F-score of 87% and an Automated Accuracy of 86%. While other models may perform better in the Empty Score metric, they do not achieve the same level of Automated Accuracy and BERT Scores as Llama 2 7b Q8. Given that we prioritize the ability of a model to extract information over empty identification, Llama 2 stands out as the best choice.

3.3.2 Human Evaluation of the best LLM

The evaluation metrics of the Llama 2 7b Q8 model are presented in Table 3.3 including the "Human Evaluation" metric, for each of the 23 questions. It is worth noting that the Empty ratio indicates the number of empty answers that were correctly predicted compared to the total dataset. For example, question 4 has an Empty ratio of 24/70 indicating that out of the 70 scientific publications that did not have any answer to question 4, Llama 2 7b Q8 was able to correctly predict an empty answer 24 times.

One notable factor contributing to the difference between the Automated Accuracy and Human Evaluation for Llama 2 7b Q8 in the scoring is the variation in their scoring calculation. The Automated Accuracy metric does not consider "empty" answers during the evaluation, whereas the Human Evaluation takes all types of answers into consideration, including the empty ones. Therefore, the Human Evaluation provides a more comprehensive representation of the model’s general behavior for each question.

Regarding the individual assessment of each question for the Human Evaluation, we can discern which questions are easier to be answered and which are more challenging. For example, question 1, a straightforward question regarding the date of the contamination event, is correctly answered 94% of the time. On the other hand, question 8, which asks about the number of people exposed to the contamination), is correctly predicted only 34% of the time. It is considered a complex question because it requires the model to accurately distinguish between exposure, illness, or number of cases.

Additionally, the difficulty of predicting the absence of an answer varies depending on the nature of the question. For example, question 11 (number of fatalities) is considered a straightforward question and that is why it is correctly predicted as empty 100 times out of 138. However, questions that are ambiguous or open to interpretation even for a human, often lead the model to hallucinate a response when it should not. Such a question is number 17, related to the case definition of the event, being correctly predicted as empty only 3 out of 84 times. The inherent complexity and ambiguity of such questions make it challenging for the model to correctly provide an answer.

Table 3.3. The results of the Human Evaluation (along with all the other metrics) for each of the 23 questions while evaluating Llama 2 7b.

Question	Precision	Recall	F1	Empty	Empty ratio	Automated accuracy	Human evaluation
1	0.89	0.86	0.88	0	0/1	0.92	0.94
2	0.90	0.90	0.90	-	-	0.83	0.83
3	0.89	0.86	0.87	-	-	0.91	0.94
4	0.89	0.89	0.89	0.34	24/70	0.80	0.66
5	0.87	0.84	0.85	0	0/11	0.7	0.59
6	0.89	0.85	0.87	0	0/28	0.76	0.52
7	0.89	0.86	0.88	0.16	4/25	0.89	0.78
8	0.84	0.82	0.83	0.15	18/116	0.66	0.34
9	0.83	0.82	0.82	0.14	2/14	0.86	0.82
10	0.88	0.83	0.85	0.11	10/87	0.91	0.52
11	0.86	0.82	0.84	0.72	100/138	0.72	0.87
12	0.89	0.85	0.87	0	0/11	0.88	0.87
13	0.9	0.84	0.87	0.51	43/84	0.94	0.76
14	0.92	0.89	0.9	0	0/3	0.96	0.97
15	0.87	0.83	0.85	0.33	1/3	0.87	0.73
16	0.88	0.84	0.86	0.11	1/9	0.91	0.87
17	0.91	0.88	0.90	0.03	3/84	0.86	0.49
18	0.9	0.85	0.87	0.03	3/80	0.90	0.52
19	0.87	0.83	0.85	0.03	1/27	0.85	0.70
20	0.87	0.87	0.87	0.02	1/46	0.79	0.65
21	0.89	0.85	0.87	0	0/58	0.95	0.66
22	0.9	0.86	0.88	0	0/81	0.91	0.53
23	0.9	0.86	0.88	0.02	2/86	0.92	0.48
Average	0.88	0.85	0.87	0.13	-	0.86	0.7

3.3.2.1 Observations in Human Evaluation

During the Human Evaluation, two key observations became quickly apparent. First, as discussed in the previous section, some questions were difficult even for a human to provide a proper answer. For example, in the following text (both sentences are from the same scientific publication) it is challenging for the annotator to decide what is considered as source of contamination (question 5): "*Epidemiological and environmental evidence suggested mains repair as the source of contamination. The wide geographical distribution of cases and preliminary interviews with patients suggested drinking water as the source of the outbreak.*" As can be seen, both main repairs and the drinking water are mentioned as sources in the text. Based on the annotation guidelines (Chapter 3.2.3), the annotator prioritizes annotating based on what humans would perceive as correct (the source of contamination of drinking water cannot be the drinking water itself), rather than trying to anticipate what the model might interpret as correct, which can lead to inconsistencies with the predicted answer of the model. For instance, the annotator might decide that the mains repair is the source, while the model predicts that drinking water is the source.

Second, there were instances where the model would provide an answer that was better than the annotated response given by the human annotator (refer to Chapter 3.2.3). For example, in response to question 3, *"Can you provide a summary of the contamination event?"*, the annotator's answer was: *"A large outbreak of cryptosporidiosis affected 43 cases in the community of Talent, Oregon. The outbreak was associated with the drinking water system in the area."*. In contrast, the model provided a more detailed response: *"The contamination event was an outbreak of cryptosporidiosis that occurred in Talent, Oregon from January to June 1992. The outbreak was associated with mechanical and operational deficiencies at one of the city's water filtration plants, which resulted in poor raw-water supply conditions and 43 cases of cryptosporidiosis. Correction of the deficiencies in the filter plant led to substantial improvement in treated water quality, which coincided with the end of the outbreak."* As can be seen, the model was able to provide a more comprehensive answer not only summarizing the incident, but also including key details and outcomes. Another example can be seen in question 15, "*What were the initial actions taken to investigate or inspect the event?"* The annotator's response was*: "1-l sample was collected from the pumping station".* In contrast, the model's response was: *"Collected three samples from the water distribution system, one sample from the tank, and one sample from the pumping station on the first day after the outbreak had been recognized (11 October 1996)."* Here, the model not only provides a more detailed response regarding the number and location of samples collected, but also includes a specific timeline, enhancing the response with temporal context.

Third, in some questions, there is a notable discrepancy between Automated Accuracy and Human Evaluation, which at first seems unusual. For example, question 17 has a high Automated Accuracy score (86%) but a relatively low Human Evaluation score (49%) across the 188 publications. This major difference can be explained by examining the Empty score and specifically, the Empty Ratio. Question 17 has a low Empty score of 3% (with an Empty Ratio of 3/84). Since Automated Accuracy does not take into account the Empty score, while Human Evaluation does, the latter has a lower score than the former.

3.3.2.2 Performance analysis across topics of questions

Figure 3.4 provides an overview of how each of the 23 questions (along with the 4 main topics of questions) scored during the Human Evaluation. In other words, how easy (or difficult) was for the Llama 2 7b Q8 model to answer the 23 questions in each of the 188 scientific publications.

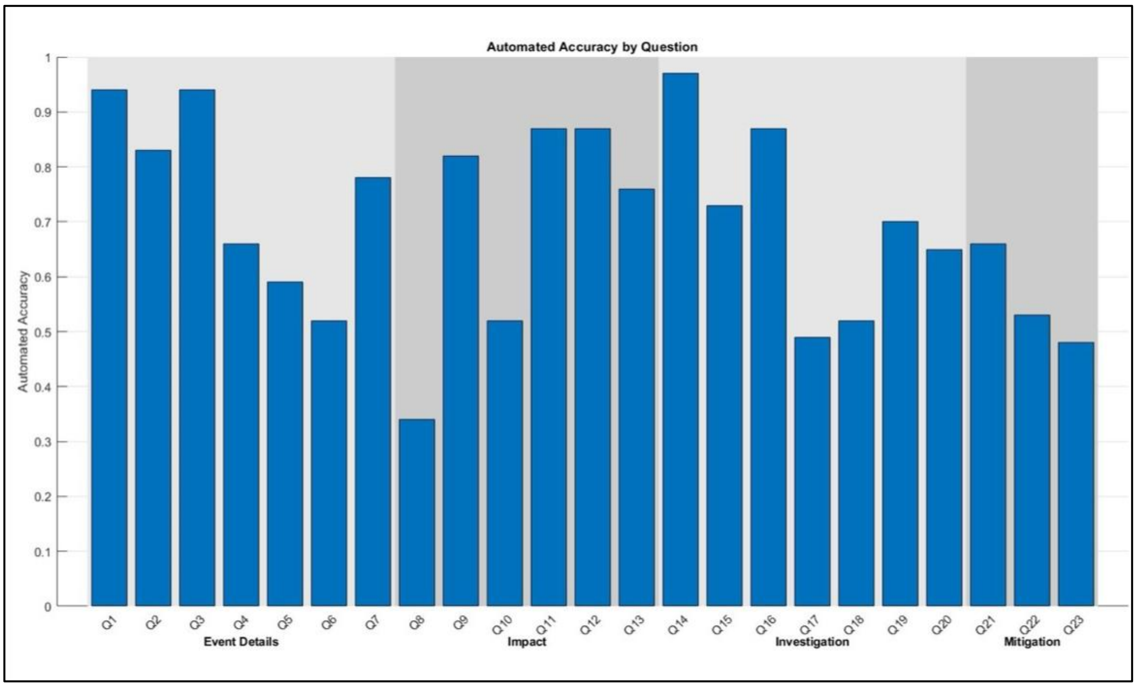


Figure 3.4. An overview of the Human Evaluation score of the 23 questions over the 188 scientific publications.

In the Event Details topic, questions 4, 5, and 6 score very low. These questions are related to the cause, the source, and the first detection of a contamination event. Therefore, a low score is somewhat expected because it is difficult for the model to distinguish between the cause and the source of contamination, or to understand the nuances between causality that leads to contamination detection versus increases in hospitalizations that trigger contamination detection.

In the impact topic, questions 8 and 10 yield the lowest score. These questions are related to the exposure and attack rate. Again, here the model struggled to differentiate between exposure, illness, and sometimes even fatalities. As for the attack rate, it is an inherently complex concept that sometimes is challenging even for humans to extract from the publication.

Regarding the investigation topic, questions 17 and 18 scored close to 50%. These questions are related to case definitions and risk factors. These questions were challenging for the model because it was not always clear from the text whether a case definition was explicitly provided. Furthermore, risk factors were often not clearly labeled as such in the text, leading the model to hallucinate and generate answers that, while logically correct based on the content, were not explicitly stated in the provided text. As a result, the model's answers were considered wrong, even though they were contextually plausible. The same applies to questions 22 and 23 from the mitigation topic. Those questions were related to the restoration and prevention measures taken by responsible authorities. It was difficult for the model to determine whether a measure was a restoration or a prevention measure, leading sometimes to the same answer for both questions.

3.4 Discussion and Recommendations

In this study, certain metrics were prioritized over others while comparing the performance of the models. The predicting ability of providing information was the primary focus (both the BERT Score and the Automated Accuracy), as it allowed us to select the model with the greatest ability to predict the correct answer for each question. However, metrics such as the Empty Score are also of great importance, as they reflect the model’s tendency to hallucinate or generate incorrect information.

This behavior is evident in the performance of the various models. The Gemma models are effective at avoiding hallucinations by providing an empty response when no answer is evident in the text, which is desirable in uncertain situations. Other studies have also found Gemma to be a robust model for minimizing erroneous outputs (Nadeau et al., 2024; Li et al., 2024a). However, their overall Automated Accuracy is notably low, having incorrect predictions on answering questions 40-50% of the time. On the other hand, the Llama model demonstrates higher Automated Accuracy in its predictions but is more prone to hallucination, often generating plausible yet incorrect answers when no clear answer is present in the text. In this study, the Llama 2 7b Q8 model was selected as the best model (having an average Automated Accuracy of 86%) primarily because the objective was to be able to extract all types of answers from the scientific publications. It was preferable to incorrectly predict one category (Empty Score metric) rather than have a model with generally low predicting ability on answering questions. However, it's important to note that in other cases, such as those requiring robust answers when the answers are not present within the provided text, the Gemma models might be a better fit, depending on the specific task requirements.

The Human Evaluation metric showed that the Llama 2 7b Q8 can provide correct answers when the question (prompt) is clear, straightforward, and easy to identify in the text (e.g., “there were 5 deaths”). However, when the question requires further interpretation or when it is relatively difficult to identify the answer in the text, the model often did not provide a correct answer. This challenge is also addressed in Nadeau et al. (2024), which examines the hallucination tendencies of Llama2, Mistral, and Gemma, finding that Llama2 tends to hallucinate more in comparison to the other LLMs, making it more difficult to control its behavior with prompt engineering. This underscores the importance of proper prompting and highlights the need for high-quality, well-structured, clear, and consistent information in the source text to ensure reliable information extraction during QA tasks.

At present, there are two main approaches to using LLMs: deploying open-source models on local systems, which typically have lower capacity, or using cloud services that offer larger, more computationally intensive models on a pay-per-use basis. In our study, we chose the first option, utilizing models with lower capacity but significantly reduced costs. Open-source models like Llama 2 offer practical, cost-effective, and scalable solutions for domain-specific tasks. However, their performance is comparable to more advanced models like OpenAI GPT (e.g., ChatGPT). Proprietary models like ChatGPT are known for their excellent reasoning capabilities, which stem from extensive training and scalability. This makes them particularly effective at handling complex queries or generating detailed insights. However, their reliance on subscription-based APIs and higher operational costs can make them less accessible, especially for organizations with limited budgets or those focusing on highly specialized tasks. Ultimately, this creates a trade-off: while models like ChatGPT are more flexible and capable of advanced reasoning, open-source alternatives can be fine-tuned for specific applications, such as answering questions related contamination events of drinking water. Furthermore, the potential in developing open-source hybrid models that combine the strengths of multiple architectures, offers high accuracy and adaptability for emergency decision support systems.

3.5 Conclusions

This study introduced an automatic information extraction approach through a QA task solution that diverges from traditional NLP techniques and the recently popular encoding-based transformers, such as BERT. The objective was to assess the performance of various LLMs on a QA task related to pathogen contamination events in drinking water. The results demonstrate that solutions based on decoder-based transformers have significantly advanced the current state of the art in information extraction.

Among the models evaluated, the Llama 2 model emerged as the best performer for general information extraction tasks. However, its inability to reliably predict empty answers -when no relevant information is present in a document- highlights an area for further research and refinement. In contrast, the Gemma model performed exceptionally well at identifying empty registers but exhibited frequent wrong predictions in other information extraction tasks. This underscores the need for complementary approaches or hybrid models that can combine strengths from both models to ensure robust performance across different dimensions of the task.

These advancements in automated extraction of information show potential for time-critical applications, such as processing large volumes of (historical) data in real-time thereby reducing the time required for manual literature reviews in case of emergencies. Therefore, further development of hybrid models could lead to robust, high-accuracy tools for emergency management and decision-support in the drinking water sector.

Acknowledgments

This research was funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 883484 (PathoCERT).

Supplementary material

Search engine	Search terms	Searched fields	Number of publications	Number of new publications
PubMed	drinking water AND outbreak AND norovirus	All fields	150	46
	drinking water AND outbreak AND Campylobacter	All fields	190	23
	outbreak AND waterborne AND E coli	All fields	384	14

Figure S3. 1 The full list of the search query for the systematic review.

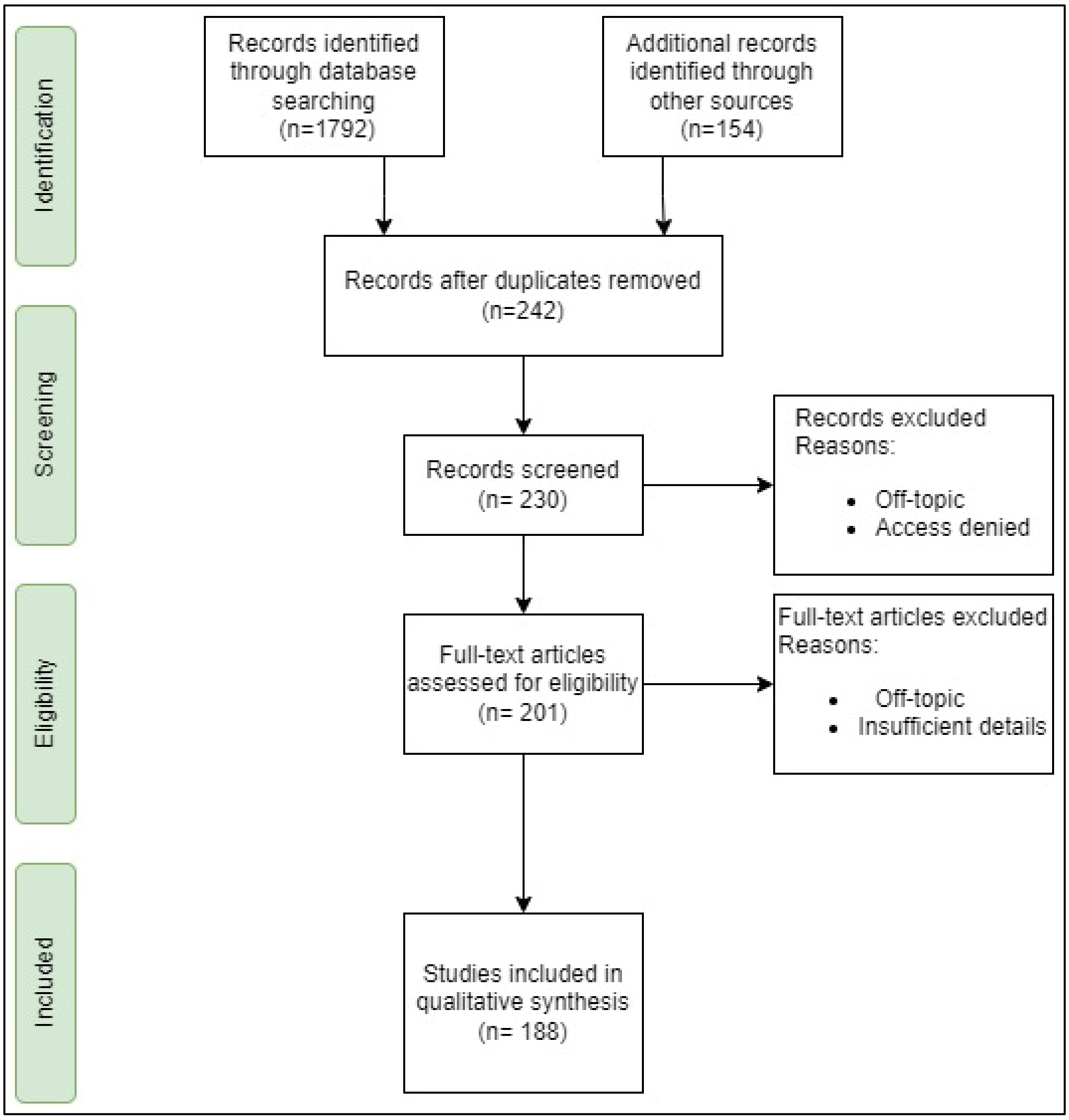
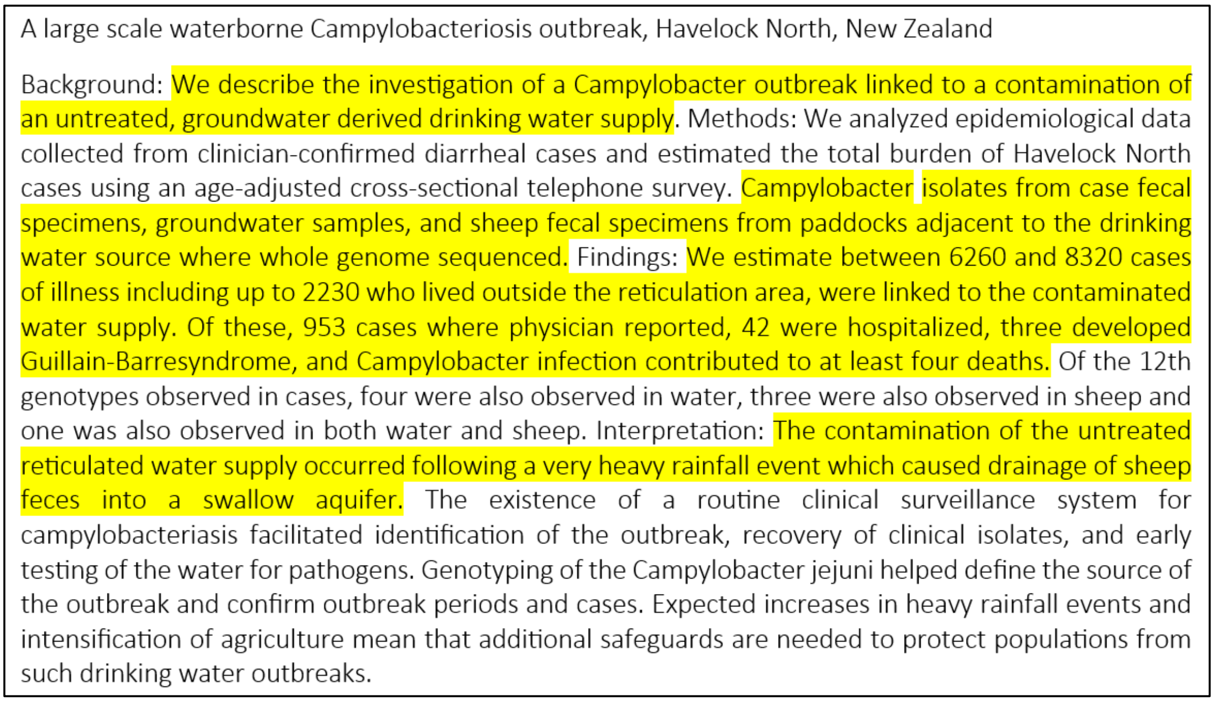


Figure S3. 2 The flowchart of the PRISMA guidelines



A large scale waterborne Campylobacteriosis outbreak, Havelock North, New Zealand

Background: We describe the investigation of a Campylobacter outbreak linked to a contamination of an untreated, groundwater derived drinking water supply. Methods: We analyzed epidemiological data collected from clinician-confirmed diarrheal cases and estimated the total burden of Havelock North cases using an age-adjusted cross-sectional telephone survey. Campylobacter isolates from case fecal specimens, groundwater samples, and sheep fecal specimens from paddocks adjacent to the drinking water source where whole genome sequenced. Findings: We estimate between 6260 and 8320 cases of illness including up to 2230 who lived outside the reticulation area, were linked to the contaminated water supply. Of these, 953 cases where physician reported, 42 were hospitalized, three developed Guillain-Barresyndrome, and Campylobacter infection contributed to at least four deaths. Of the 12th genotypes observed in cases, four were also observed in water, three were also observed in sheep and one was also observed in both water and sheep. Interpretation: The contamination of the untreated reticulated water supply occurred following a very heavy rainfall event which caused drainage of sheep feces into a swallow aquifer. The existence of a routine clinical surveillance system for campylobacteriasis facilitated identification of the outbreak, recovery of clinical isolates, and early testing of the water for pathogens. Genotyping of the Campylobacter jejuni helped define the source of the outbreak and confirm outbreak periods and cases. Expected increases in heavy rainfall events and intensification of agriculture mean that additional safeguards are needed to protect populations from such drinking water outbreaks.

Figure S3. 3 An example of the data processing step highlighting the relevant information within the introduction section of a publication.

References

Anand, C., & Apul, D. (2011). Economic and environmental analysis of standard, high efficiency, rainwater flushed, and composting toilets. Journal of Environmental Management, 92(3), 419–428.

Chen, Q., Du, J., Hu, Y., Keloth, V. K., Peng, X., Raja, K., ... & Xu, H. (2023). Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv preprint arXiv:2305.16326.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hrudey, S. E., & Hrudey, E. J. (2004). *Safe drinking water*. IWA publishing.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kartchner, D., Ramalingam, S., Al-Hussaini, I., Kronick, O., & Mitchell, C. (2023, July). Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models. In The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks (pp. 396-405).

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

Li, H., Chi, H., Liu, M., & Yang, W. (2024a). Look Within, Why LLMs Hallucinate: A Causal Perspective. *arXiv preprint arXiv:2407.10153*.

Li, H., Liu, J., Wang, Z., Luo, S., Jia, X., & Yao, H. (2024b). LITE: Modeling Environmental Ecosystems with Multimodal Large Language Models. arXiv preprint arXiv:2404.01165.

Liang, J., Yu, X., Hong, W., & Cai, Y. (2024). Information extraction of UV-NIR spectral data in waste water based on Large Language Model. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 124475.

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

Martin, S., Penttinen, P., Hedin, G., Ljungström, M., Allestam, G., Andersson, Y., & Giesecke, J. (2006). A case–cohort study to investigate concomitant waterborne outbreaks of Campylobacter and gastroenteritis in Söderhamn, Sweden, 2002–3. *Journal of water and health*, *4*(4), 417-424.

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

Nadeau, D., Kroutikov, M., McNeil, K., & Baribeau, S. (2024). Benchmarking Llama2, Mistral, Gemma and GPT for Factuality, Toxicity, Bias and Propensity for Hallucinations. *arXiv preprint arXiv:2404.09785*.

Paraskevopoulos, S., Vrachimis, S., Kyriakou, M., Pavlou, P., Kouzapas, D., Milis, G., Smeets, P., Eliades, D., Medema, G., Polycarpou, M., & Panayiotou, C. (2022). PathoINVEST: Pathogen contamination investigations during emergencies. In 2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Indsustry. Universitat Politecnica de Valencia. https://doi.org/10.4995/WDSA-CCWI2022.2022.14799

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

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ribalta, M. (2024). *PathoTHREAT-data-retrieval* [Source code]. GitHub.https://github.com/Applied-Artificial-Intelligence-Eurecat/PathoTHREAT-data-retrieval

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chapter 4

Modeling the health impact of wastewater contamination events in drinking water networks

Published as original research article in the Journal of Cleaner Production, 143997.
Paraskevopoulos, S., Vrachimis, S., Kyriakou, M., Eliades, D. G., Smeets, P., Polycarpou, M., & Medema, G. (2024). https://doi.org/10.1016/j.jclepro.2024.143997

Chapter 4- Modeling the health impact of wastewater contamination events in drinking water networks.

Abstract

Pathogen intrusion in drinking water systems can pose severe health risks. To better prepare in planning and responding to such events, computational models that capture the intrusion and health impact dynamics are needed. This study presents a novel benchmark testbed that integrates current knowledge on pathogen transport and fate in chlorinated systems and can assess infection risk from contamination events. The model considers organic matter degradation, chlorine decay mechanisms, pathogen inactivation kinetics, as well as stochastic water demands.

We studied modeling of wastewater intrusion events that can occur anywhere within a chlorinated and non-chlorinated network. We applied the Quantitative Microbial Risk Assessment framework focusing on three pathogens: enterovirus, *Campylobacter*, and *Cryptosporidium*, and their respective dose-response models. Synthetic household-level water demand time series were used to model the individual water consumption timing and calculate the infection risk (exposure via ingestion).

Model outcomes indicate that while chlorination aids mitigation, larger contaminations can still lead to infections due to chlorine resistance (for Cryptosporidium) and chlorine depletion at the contamination point. In our example scenarios, chlorine-susceptible pathogens infected 0.78-26.6 % of the downstream population, while chlorine-resistant ones infected the entire downstream population. Enterovirus infection risk is higher, despite the concentrations in the contamination source being lower, due to the lower susceptibility to chlorine than Campylobacter. In non-chlorinated networks, the modeled wastewater contamination events led to 11-46 % infection risk in the total population, depending on the contamination location. Hydraulic uncertainty had a limited influence on infection risk. Furthermore, Campylobacter's infection risk is more sensitive to the initial concentration in the contamination source whereas enterovirus infection risk to the inactivation rate. The model further indicates that the time window for effective mitigation of the magnitude of a waterborne outbreak is short (within hours).

Keywords: QMRA; Wastewater contamination; Drinking water network modeling; Pathogens; EPANET-MSX

4.1 Introduction

Safe drinking water is crucial for society, impacting health and well-being. Drinking water distribution networks (DWDN) are critical infrastructures, recognized by USA's Presidential Policy Directive 21 and the European Union's Directive (EU) 2022/2557. This requires plans to enhance water suppliers' resilience against natural, accidental, and malicious threats (Teixeira et al., 2019). One such threat is when wastewater intrudes into the DWDN. This can expose thousands to contaminated tap water, causing acute health effects from pathogens (Hrudey and Hrudey, 2004). Proper DWDN operation and maintenance ensures hygiene and hydraulic integrity, preventing pathogen intrusion (Medema et al., 2013). In systems with residual disinfectants, high disinfectant concentration can maintain safety when standard conditions are not met (Lechevallier, 1999). However, outbreaks have been linked to fecal contamination in (chlorinated) networks, often due to cross-connection between sewage- and drinking-water pipelines or intrusion during main breaks (Craun and Calderon, 2001; Hrudey and Hrudey, 2019). Specifically, van Lieverloo et al. (2007) note that 26% of the outbreaks in the UK from 1911 to 1995 were caused by failures in the DWDN. Additionally, 18-20% of outbreaks in Nordic countries from 1975 to 1991 and 18% of outbreaks in the USA from 1971 to 1998 were attributed to similar failures. Notable incidents such as the wastewater contamination in Nokia, Finland with 8453 cases (Laine et al., 2011), or the wastewater infiltration events in Italy (Giammanco et al., 2018) and in Denmark (Kuhn et al., 2017) resulting in 25 and 63 cases respectively, highlight the vulnerability of those systems to contamination. During a contamination event, authorities must assess health impacts and respond quickly and effectively. Accurate representation of contamination type and site, demand-driven hydraulics, understanding of contaminant transport, and the effect of a residual disinfectant as a mitigation are crucial for a comprehensive assessment and rapid, efficient response.

An approach to assess the risk during contamination in the DWDN is to perform a Quantitative Microbial Risk Assessment (QMRA). This estimates customers' exposure to enteric pathogens through ingestion. Studies have combined hydraulic modeling and QMRA for evaluating health risks after wastewater intrusion in DWDNs. For instance, Teunis et al. (2010) examined the risk of norovirus intrusion from sewers into DWDNs due to negative pressure transients. They used the EPANET-MSX hydraulic model and Monte Carlo simulations for random virus entry and dilution estimation. Their study considered the coincidence of virus presence and tap water usage, finding that this factor significantly affects the calculated infection risk level and distribution in the population. Another effort described by Yang et al. (2011) involved surge modeling and hydraulic simulations to model a virus intrusion in a DWN again due to pressure transients. The authors employed EPANET-MSX to integrate a Chick-Watson model that accounted for the inactivation kinetics of selected pathogens and chlorine decay. They concluded that the factors influencing the risk of viral infection were the duration of the negative pressure event and the number of affected nodes, without incorporating stochastic water demand or other water quality parameters. Blokker et al. (2018) developed a QMRA model for contamination events after main repairs in non-chlorinated DWDNs. They discovered that pathogen concentration greatly influences the ingested dose and that the infection risk varies notably between pathogens due to different dose-response relationships.

Controlling pathogens in the DWDN heavily relies on residual disinfectants. Over the past 25 years, there's been increasing interest in modeling chlorine (Cl) transport and decay due to reactions with total organic carbon (TOC) in DWDNs. Frankel et al. (2023) assessed the uncertainty of drinking water quality in DWDNs, specifically focusing on monochloramine decay. They focused on quantifying the effects of hydraulic and chemical uncertainties on water quality predictions. They conducted a sensitivity analysis and Monte Carlo simulations to identify the most influential chemical parameters and explore the impact of both chemical and hydraulic uncertainty. Their findings were that monochloramine uncertainty is significantly influenced by hydraulic variability and increases as water age increases. Their study emphasized the importance of accounting for these uncertainties to make accurate model-based decisions for managing water quality in DWDNs. Pelekanos et al. (2021) applied a parallel first-order bulk and wall chlorine decay model to evaluate a network's vulnerability to deliberate contamination attacks, using nominal water demands. They found that contamination location significantly affects the size of the exposed population. Abhijith et al. (2021) examined a chlorinated (and chloraminated) network's response to arsenic contamination using second-order kinetics, based on competing reactions in water. This study emphasized the critical role of disinfectant residual. Lastly, Fisher et al. (2017a) used a two-reactant model with fast and slow agents, incorporating temperature as it greatly affects bulk chlorine decay.

Eliades et al. (2023) provided a detailed review of contamination event diagnosis tools, emphasizing the need for realistic physical and virtual testbeds to simulate contamination emergencies and assess their impact, considering uncertainties. However, most models use generic contamination approaches, lacking tools to accurately represent pathogen dynamics- and failing to consider all important modeling parameters together. To the best of the author's knowledge, there has not been any other attempt to model all the different reactions that occur in a DWDN simultaneously during a wastewater contamination event, while also using stochastic water demands and assessing the infection risk using QMRA.

In line with this, to enable responsible authorities to prepare for and respond effectively to contaminations, we have developed a novel, open benchmark testbed named BeWaRE (Benchmark for Water network and Risk Evaluation). This testbed integrates all current relevant knowledge regarding the transport and fate of pathogens in chlorinated systems in one model and is capable of estimating the health impacts of such events. With BeWaRE, responsible authorities can model various contamination events. These events vary from minor, where negative pressure transients cause slight wastewater entry, to major accidents with significant wastewater influx. The benchmark can be used for developing new software and decision support tools for monitoring, control, and management of contamination emergencies, as well as creating datasets for machine learning research.

BeWaRE integrates the findings of previous studies, accounting for bulk and wall chlorine decay, various pathogen inactivation kinetics, TOC degradation, and realistic water demands and consumption distribution, incorporating QMRA.

The contributions of this work are summarized below:

- Introduction of an open-access testbed designed for comprehensive simulation of contamination by waterborne pathogens under various disinfection regimes.
- Integration of realistic household consumption profiles improving daily consumption pattern accuracy and health impact calculations for waterborne pathogens using QMRA.
- Investigation of a wastewater contamination event and evaluation of the importance of the input parameters.

Specifically, this Chapter presents a benchmark hydraulic and water-quality model to assess the health impact following a large wastewater contamination in a chlorinated and non-chlorinated network and evaluate chlorine's mitigating effect. The model was tested using a modified version of L-Town, a benchmark network from the BattLeDIM (Battle of the Leakage Detection and Isolation Methods) competition (Vrachimis et al., 2022), featuring 782 junctions, 905 pipe segments, and serving approximately 28,000 citizens. In our example, we investigated a single contamination event originating from three distinct locations using *Campylobacter*, enterovirus, and *Cryptosporidium* as reference pathogens. Hydraulic uncertainty was addressed to account for the dynamic, uncertain nature of water demands and examine their influence on the model outcome. Water quality uncertainty was included to examine the effects of variability of input parameters on the model outcome. The model outcome is the expected health impact over time and space, expressed as the expected number of infections, and the infection risk, following the QMRA steps.

4.2 Benchmark model development

The benchmark model developed is depicted in Figure 4.1. The hydraulic modeling includes stochastic water demands that determine the hydraulics of the entire network. This involves modeling different water end-uses, from which we isolate the tap water end-use. As shown in the figure, we generate individual tap water consumption events that will later be used for the risk assessment. The hydraulic modeling is also integrated with the water quality modeling component. The water quality modeling includes simulating reactions between different agents of interest, eventually leading to the calculation of pathogen concentration. This concentration, combined with the volume calculated from the consumption events, is used to determine the dose. The dose is then integrated into the QMRA part to assess the infection risk. All the different components of the BeWaRE model are described in detail in the following sections.

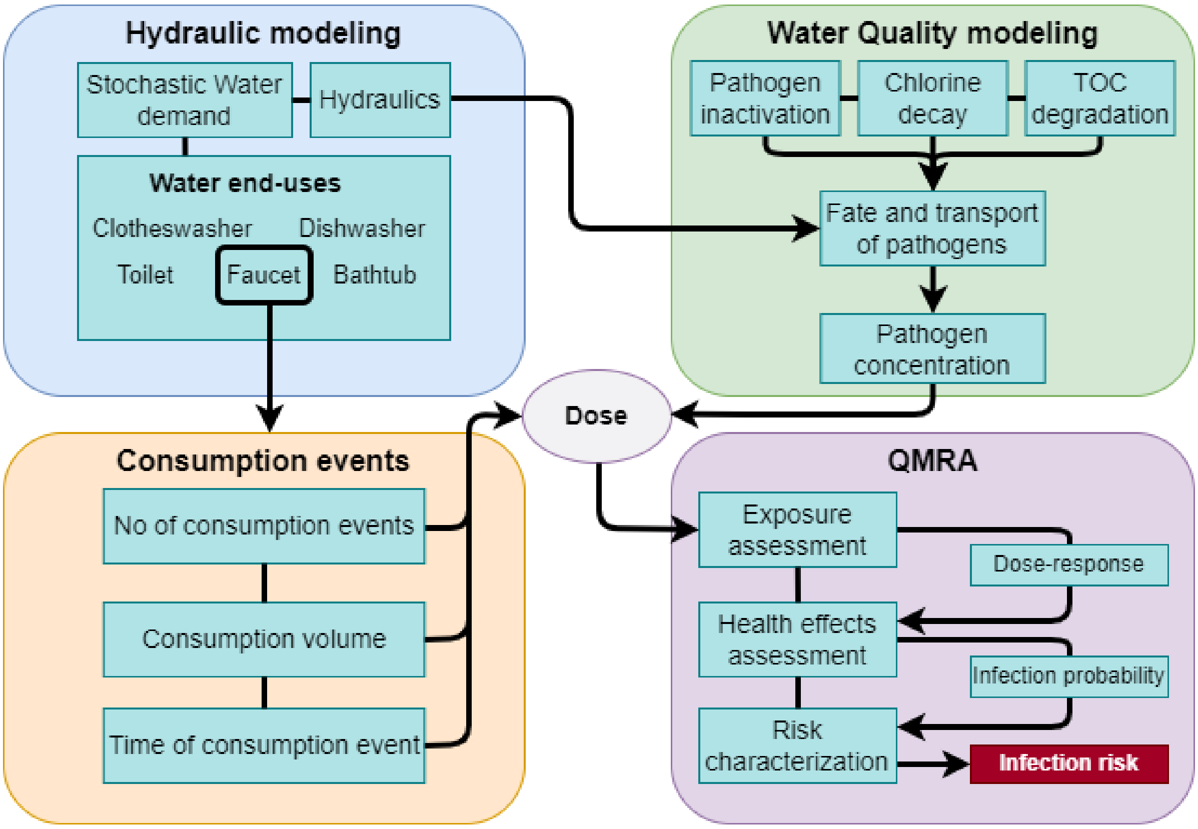


Figure 4. 1 The flowchart of BeWaRE model with the integration of all components.

4.2.1 Network graph

The topology of the DWDN is modeled by a directed graph denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V}$ is the set of nodes such that $\mathcal{V} \subset Z \times \Theta_v$. The set $Z = \{1, \cdots, n_v\}$ indicates the positive integers representing the index of the i-th node, $v_i \in \mathcal{V}$, and $|\mathcal{V}| = n_v$ is the total number of nodes. Nodes represent pipe junctions and consumer (water demand) locations, reservoirs, and tanks. The set Θ_v associates each node with parameters (real numbers) detailing the network's physical properties that affect water flow and quality, such as node elevation. Each node v_i is associated with a time-varying consumer water demand, denoted by $d_i(t)$. The set $\mathcal{E}$ represents edges (links) defined as $\mathcal{E} \subset \mathcal{V} \times \mathcal{V} \times \Theta_e$. An edge $e_{(i,j)} \in \mathcal{E}$ connects nodes v_i and v_j where $i, j \in Z$ and $i \neq j$. The total link count is $|\mathcal{E}| = n_e$. Links represent pipes, pumps, and valves, with pumps and valves being the main hydraulic control elements in a DWDN. The set Θ_e associates the edge with its parameters (real numbers). Depending on the edge type, parameters vary. For instance, a pipe might have length, diameter, and roughness as parameters, while a pump's parameters might be polynomial coefficients defining its characteristic curve. In this work, both Θ_v and Θ_e parameters are considered time-invariant since they refer to characteristics of pipes and nodes that may change slowly over time (e.g., years), while we consider wastewater contamination events lasting from hours to days.

4.2.2 Consumer demand modeling

The main driver of water network hydraulics is consumer demand at nodes (demand-driven analysis), represented as $d_i(t)$ for the time-varying demand at node v_i. Typically, demand is modeled by an average or base demand component, multiplied by a daily or weekly consumption pattern. Approximations of base demand data can be deduced from the utility's billing records, while patterns are usually rough approximations (Vrachimis et al., 2019). We employ the *STochastic Residential water End-use Model* (STREaM) tool, to generate synthetic household-level water demand time series. STREaM uses a large dataset with observed and disaggregated water end-uses from over 300 single-family U.S households (Cominola et al., 2018). The associated water end-uses are toilet, faucet, bathtub, clothes washer, and dishwasher. Each water end-use has distinct consumption patterns and probability distributions for water use volume, use duration, daily frequency, and time of use during the day. This gives a realistic residential demand profile for the L-Town network, ignoring non-residential demands such as industries. To derive a daily demand time-series $d_i(t)$ per node, we use the L-Town network's base demand to calculate the population associated to node v_i, defined as $Pop_{(i)}$, assuming an average consumption of 150 L/day per person. Multiple simulations of the STREaM tool are employed with different household occupancy to allow for variations in consumption patterns until the total occupancy equals $Pop_{(i)}$. The demand profile for node v_i, indicated by $d_i(t)$, is the combined household consumption at each time instant. Note that, using this demand modeling approach, different daily tap water end-uses per individual can be distinguished at each node and used in the exposure assessment (part of QMRA as discussed in Chapter 1.2.2.2).

4.2.3 Hydraulic dynamics

The key hydraulic quantity associated with each node v_j is the *hydraulic head*, denoted by h_j. The main hydraulic quantity associated with a link $e_{(i,j)}$ is the *water flow*, denoted by $q_{(i,j)}$ (Boulos et al., 2006). The overall hydraulic state $x^h \in R^{n_h}$ of a DWDN is defined by the head at nodes and flow in links, thus $n_h = n_v + n_e$. These states are calculated using a hydraulic model of a DWDN, which is a set of equations derived from the laws of (i) conservation of mass; and (ii) conservation of energy in the network. In this work, we use the EPANET modeling software (Rossman, 2000) to solve these equations, which uses the *pipe formulation* as proposed by Todini (1987).

4.2.4 Water-quality dynamics

Water quality characterizes the concentration of key variables in water, while water quality dynamics describe changes in the concentration of physical, chemical, and biological agents within the DWDN over time and space. Agents in DWDN either react with others altering concentration over time or maintaining a constant concentration. Both types of agents are diluted and transferred within the water, thus their concentration at a particular network location change over time.

Let $W \in R^{n_w}$ be a vector indicating the concentration of n_w agents of interest in a DWDN, at a certain location and time, with $w^{(i)}$ being the i-th agent. We focus on certain agents because they either need to be controlled or because they may react with these controlled agents.

Reaction dynamics explain how agent concentrations change due to reactions or decay. Single-species reaction dynamics, commonly used in water quality modeling literature, describe the decay rate of an agent (Clark et al., 2010), representing the concentration of a single agent, $w^{(i)}$, while neglecting others. This simplification is convenient since we typically don't know all reactions and agents present in water.

During normal operation, hydraulic dynamics in a water network influence the water quality dynamics through agent transport along pipes and dilution at pipe junctions. The change in agent concentration over time t and space, coupled with reaction dynamics in bulk water and on pipe surfaces (axial dispersion neglected for simplicity), is represented by a first-order hyperbolic partial differential equation (Eliades et al., 2023):

$$\frac{\partial W_{(i,j)}(z,t)}{\partial t} + \frac{q_{(i,j)}(t)}{\alpha_{(i,j)}} \frac{\partial W_{(i,j)}(z,t)}{\partial z} = f_r\big(W_{(i,j)}(z,t), \Theta_r\big) + B(z)u_{(i,j)}(z,t) + B(z)\phi_{(i,j)}(z,t) \quad 1$$

where $W_{(i,j)}(z,t)$ is the agent concentrations vector in water at continuous time t and at distance z along a pipe corresponding to the edge $e_{(i,j)}$, with water flow $q_{(i,j)}(t)$ and pipe cross-sectional area $\alpha_{(i,j)} \in \Theta_e$. The function $f_r(\cdot)$ denotes concentration changes due to reactions with other agents in the water or on pipe walls, considering the pipe parameter vector Θ_r. The function $u_{(i,j)} \in R^{n_w}$ represents controlled agent input (e.g., disinfectant addition). The function $\phi_{(i,j)}(z,t) \in R^{n_w}$ corresponds to the uncontrolled injection of contaminants that can occur at any location in a DWDN (e.g., wastewater intrusion). The matrix $B(z)$ specifies the injection location and agent type. Note that, if a new agent is added to the network, this needs to be included in W, and suitably modify function $f_r(\cdot)$ if this reacts with other agents.

In general, this hyperbolic partial differential equation cannot be solved analytically, however, a numerical solution is possibly by using a suitable discretization method. One approach is to segment the network into finite volumes, and model multi-species reactions (Shang et al., 2008a) as coupled sets of differential and algebraic equations solved for each finite volume of the network, summarized by:

$$\frac{\mathrm{dW(t)}}{\mathrm{dt}} = \mathrm{f_r}(\mathrm{W(t)}, \Theta_\mathrm{r}) \qquad 2$$

$$\mathrm{f_g}\big(\mathrm{W(t)}, \Theta_\mathrm{g}\big) = 0 \qquad 3$$

where W is a vector of average concentrations of n_w agents of interest within a finite volume, $f_r(\cdot)$ is a vector field denoting concentration change due to decay reactions between agents, $f_g(\cdot)$ corresponds to the algebraic equations for mass balance, and Θ_r, Θ_g are the coefficients of the reaction kinetics.

In this work, we used EPANET for hydraulic modeling and EPANET-MSX for multi-reaction modeling, chosen for their open-source tool ecosystem, for instance, the EPANET-MATLAB Toolkit for effective scenario simulations in MATLAB (Eliades et al., 2016). Regarding advection dynamics, this benchmark model employs the EPANET-MSX simulator with the following core assumptions:

a) Advective transport in pipes: agents move with the fluid's average velocity and interact with other species and pipe walls.
b) Mixing at pipe junctions: Inflows from multiple links are assumed to mix completely and instantly.
c) Mixing in storage nodes: all inflows to tanks mix completely with existing contents, subject to possible bulk phase reactions, with alternative schemes available to model plug flow.

4.2.5 Agents of interest during wastewater intrusion

Water quality dynamics largely depend on the chosen agents W and the differential equations $f_r(\cdot)$ that describe their reactions. For example, contaminants may react with disinfectants, reducing disinfectant concentration.

Table 1 lists the reference pathogens that were modeled, each representing a pathogen group (bacterium, virus, protozoon) with varying Cl resistance and infectivity. The selection of these pathogens is primarily due to their frequent occurrence in wastewater, differences in chlorine resistance, and high infectivity. Their data availability and use in existing literature, offers a comparative and well-established basis for their inclusion in our analysis (Betanzo et al., 2008, Laine et al., 2011, Odhiambo et al., 2023). For modeling, we denote pathogens as the agents of interest, represented by $w^{(1)} = C_P$ (organisms/L). They enter drinking water during a large wastewater intrusion event, i.e., the first contaminant input $\phi^{(1)} = P$ (organisms).

Table 4. 1 Waterborne pathogens and their significance in water supplies. Adapted from WHO. Guidelines for drinking water quality (World Health Organization, 2017).

Pathogen	**Health significance**	**Persistence in water supplies**	**Chlorine resistance**	**Relative infectivity**
Campylobacter	High	Moderate	Low	Moderate
Enterovirus	High	Long	Moderate	High
Cryptosporidium	High	Long	High	High

Cl (mg) is a key agent of interest, as it impacts pathogen concentration. We denote the concentration of chlorine as $w^{(2)} \equiv C_{Cl}$ (mg/L), while the injected concentration of chlorine is a controlled input $u^{(1)} \equiv C_{Cl}$ (mg/L). Wastewater carries organic and inorganic compounds reacting with chlorine in a chlorinated network. We designated TOC as an indicator of all chlorine-reducing agents (CRA) in water that includes both Natural Organic Matter (NOM), typically considered as slow chlorine-reducing agents (SRA), and the inorganic compounds (such as ammonia and iron), that are typically fast chlorine-reducing agents (FRA), as seen in other studies (Vieira et al., 2004; Monteiro et al., 2014; Fisher et al., 2017a). The use of TOC as an indicator is convenient since it is measurable, however, it is important to note that not all TOC contributes directly to chlorine demand, since it includes both reactive and non-reactive compounds. FRA and SRA from wastewater are denoted by $w^{(3)} \equiv C_{FRA}$(mgCl-equiv/L), $w^{(4)} \equiv C_{SRA}$(mgCl-equiv/L), and modeled as contamination inputs $\phi^{(2)} \equiv FRA$ (mgCl-equiv), $\phi^{(3)} \equiv SRA$ (mgCl-equiv). CRA is found at lower levels in drinking water than in wastewater and mostly consists of SRA. We account for this by inserting additional SRA at DWDN entry points, denoted by $u^{(2)} \equiv C_{SRA}$(mgCl-equiv/L).

The complete state, control input, and contamination input vectors for this benchmark model are then given by:

$$W = [C_P, C_{Cl}, C_{FRA}, C_{SRA}]^{\top}, U = [C_{Cl}, C_{SRA}]^{\top}, \Phi = [P, FRA, SRA]^{\top} \qquad 4$$

4.2.6 Modeling reactions

The water quality dynamics cover concentrations of four agents: Chlorine C_{Cl} (mg/L), fast and slow reducing agents C_{FRA}, C_{SRA} (mgCl-equiv/L) , and various reference pathogens C_{P_i} in Colony Forming Units (CFU/L), plaque-forming units (PFU/L) and oocysts/L, all assumed viable/infectious at intrusion. Figure 4.2 shows processes in a pipe during wastewater intrusion into the DWDN. Chlorine decay is modeled in both the bulk water and near the pipe wall. In the bulk phase, chlorine reacts with SRA at DWDN entry points. From the intrusion point (in the DWDN), it reacts with both FRA and SRA, causing bulk chlorine decay. In the wall phase, chlorine reacts with the biofilm on the network pipe walls, leading to further chlorine decay. The equation for bulk and wall chlorine decay is described as:

$$\frac{dC_{Cl}}{dt} = f_{Cb}(C_{FRA}, C_{SRA}, C_{Cl}) + f_{Cw}(C_{Cl}) \qquad 5$$

where C_{Cl} is the total chlorine concentration (mg/L) at time t (time notation omitted for simplicity), $f_{C_b}(\cdot)$ and $f_{C_w}(\cdot)$ are functions describing chlorine (mg/L) reactions in the bulk and wall phase respectively. The reaction of chlorine with FRA, SRA, and pathogens P, results in the degradation of the first two, and the inactivation of the latter.

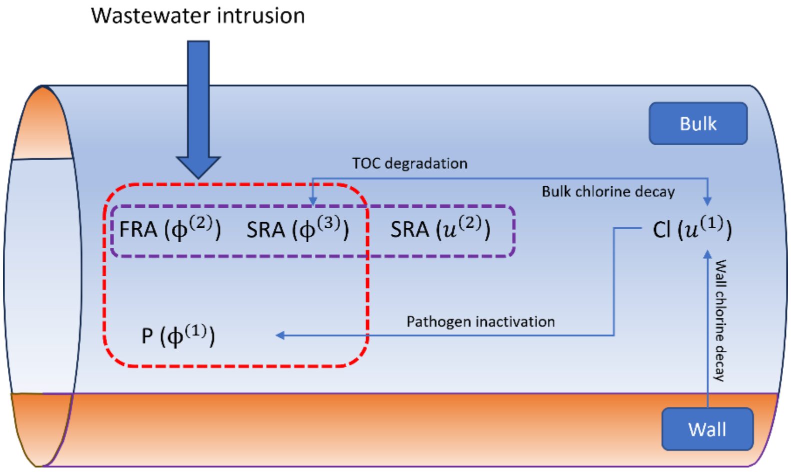


Figure 4. 2 Schematic showing processes during wastewater intrusion. P ($\phi^{(1)}$), FRA ($\phi^{(2)}$) and SRA ($\phi^{(3)}$) are the pathogens, the fast and the slow reducing agents respectively entering from wastewater intrusion. Cl ($u^{(1)}$) and SRA ($u^{(2)}$) are the controlled chlorine input and the slow reducing agents respectively entering through entry points.

The proposed model was applied on a real network, on which the L-Town benchmark was based. The parameters of chlorine decay (see following sections) were calibrated using real chlorine measurements from sensors installed in that network. Parameters that could not be validated from the calibration process were based on the literature. More details on the calibration process are provided in the following sections and in the supplementary material.

4.2.6.1 Bulk chlorine decay

The parallel second-order model is commonly used for bulk chlorine decay, accounting for fast and slow reactions with reactants (Monteiro et al., 2014; Fisher et al., 2017a). In the fast phase, chlorine reacts with inorganic compounds and highly reactive organic compounds, represented as FRA. In the slow phase, it is consumed by less reactive organic compounds, modeled as SRA. Chlorine's reactions with pathogens are insignificant compared to those with FRA and SRA, thus they're neglected. The two-phase equation is shown below:

$$f_{Cb}(C_{FRA}, C_{SRA}, C_{Cl}) = -k_{\mathrm{FRA}} C_{\mathrm{FRA}} C_{Cl} - k_{\mathrm{SRA}} C_{\mathrm{SRA}} C_{Cl} \qquad 6$$

Where k_{FRA}, k_{SRA} (L/mgh) are the decay rate coefficients for fast and slow reactions respectively.

We adopted chlorine decay parameter values from Monteiro et al. (2014) as they also examined chlorine decay in a DWDN under similar conditions of water temperature, organic material, and chlorine levels to our contamination scenario. For the value of slow reaction decay rate coefficient (k_{SRA}), we verified that the predicted values of chlorine concentration aligned with the observed sensor data. The high value of fast reaction decay rate coefficient (k_{FRA}) reported by Monteiro et al. (2014) is assumed to be suitable for our model, since high concentrations of ammonia and other inorganic compounds are expected in wastewater.

4.2.6.2 Wall chlorine decay

We followed the work of Monteiro (2020) where the authors used the EXPBIO wall decay model by Fisher et al. (2017b) to study chlorine decay from biofilm activity in a full-scale DWDN, using first-order kinetics:

$$f_{Cw}(C_{Cl}) = -\frac{4}{D}\left(\frac{Ae^{-BC_{Cl}}}{1 + Ae^{-BC_{Cl}}/(km)} C_{Cl}\right) \qquad 7$$

where D is the pipe diameter (dm), km the mass transfer coefficient (dm/h), A an amplification factor (dm/h), and finally B the rate coefficient (L/mg).

In the real network (the basis of L-Town), six chlorine sensors were installed to record chlorine concentrations at five-minute intervals. To calibrate the proposed water quality model, the network with calibrated hydraulics and known chlorine input was first simulated for one week. The wall decay parameters A and B were then adjusted to minimize the error between the model predictions and actual measurements from the chlorine sensors and ensure that the simulated chlorine residual closely resembles reality. After calibration, the parameter B was held constant at 14 (L/mg) while the A value ranged between [0.01,1], as it is related to the pipe material. Specifically, we linked A values to pipe roughness, since it is expected to correlate with the level of biofilm formation (Douterelo et al., 2016). The amount of biofilm differs across pipes, with areas that have high biofilm thickness exhibiting high chlorine demand, and areas with low thickness having lower chlorine demand. PVC pipes with roughness coefficient > 140 (Hazen-Williams), less prone to biofilm, got $A = 0.01$. Cast/galvanized iron pipes with roughness < 140, more prone to biofilm formation and wall chlorine decay, got A values between [0.01,1], with $A \in R$.

4.2.6.3 CRA degradation

A CRA concentration of 140 mg/L based on TOC concentration in wastewater was taken from (Metcalf et al., 1991), representing all chlorine-reducing agents. Based on Fisher et al. (2017a), the concentrations of chlorine-reducing agents that react fast (FRA) is approximately 40% of the total chlorine-reducing agents' concentration (expressed as mgCl-equiv/L) while SRA constitute around 60%. Similarly, in the paper of Vieira et al. (2004), chlorine decay follows a similar 40%-60% pattern, indicating that the fraction of chlorine that reacts fast is approximately 40%, while the rest is 60%. This approximation serves as a practical baseline for modeling the reactive fractions of CRA. To estimate the amount of CRA entering the DWDN at its entry points, we used the SRA concentration from Monteiro et al. (2014). This refers to the CRA naturally present in the DWDN. The formula that describes the degradation of FRA and SRA is given by Monteiro et al. (2014):

$$\frac{dC_{\mathrm{FRA}}}{dt} = -k_{\mathrm{FRA}} C_{\mathrm{FRA}} C_{Cl} \tag{9}$$

$$\frac{dC_{\mathrm{SRA}}}{dt} = -k_{\mathrm{SRA}} C_{\mathrm{SRA}} C_{Cl} \tag{10}$$

4.2.6.4 Pathogen inactivation

Pathogen inactivation by chlorine is commonly modeled as Chick Watson kinetics (Teunis et al., 2010; Betanzo et al., 2008):

$$\frac{\mathrm{d}C_P}{\mathrm{d}t} = -k_p(T) C_P C_{Cl} \tag{11}$$

where C_P is pathogen concentration (CFU or PFU or oocysts/L) at time t (time notation omitted for simplicity), $k_p(T)$ is the temperature-dependent inactivation rate (L/mg h), and T is the temperature in degrees Celcius. The inactivation rate for *Campylobacter* was taken from Betanzo et al. (2008), as they also modeled microbial intrusion in a chlorinated network. Enterovirus inactivation rates were derived from Rachmadi et al. (2020), who studied chlorine inactivation of coxsackie virus, using the rate calculated from required CT values of 4 log inactivation at 5 degrees Celsius.

Temperature dependence

Using the Arrhenius equation, we defined the pathogen inactivation rate $k_p(T)$ for enterovirus and *Campylobacter* at a given temperature:

$$k_p(T) = Ae^{(-Ea/R(T+273))} \tag{12}$$

where A is the frequency factor (L/mg h), E_a the activation energy (J/mol), and R the gas constant (J/K mol). For *Cryptosporidium*, we assumed a zero-inactivation rate due to its chlorine resilience. We used inactivation rates at two different temperatures to calculate A and E (Betanzo et al., 2008; Rachmadi et al., 2020).

4.3 Quantitative Microbial Risk Assessment

4.3.1 Exposure assessment

We assume that exposure to pathogens occurs only through the ingestion of tap water. We also assume that an individual person consumes 1 L of drinking water per day, divided into several consumption (exposure) events of 0.25 L or less. From all the end-uses generated by the stochastic demand generator, we isolated the tap water end-use (faucet), and considered the opening of the kitchen tap as the event that people drink water. Figure 4.3 illustrates an example of multiple daily consumption events by 126 individuals in a node, where each rectangle represents an individual consumption event, while the color indicates the volume of water consumed. The plot shows cumulative tap water consumption for the day. From this, the variability and distribution of tap water consumption behavior throughout the day for each individual is evident. Some individuals drink 1 L using only four consumption events, while others use six or seven consumption events. This variability reflects the different levels of exposure of each individual throughout the day. Pathogens ingested per consumption event are found by multiplying consumed water volume with pathogen concentration at each timestep. The daily dose per individual sums up the number of pathogens from all daily consumption events.

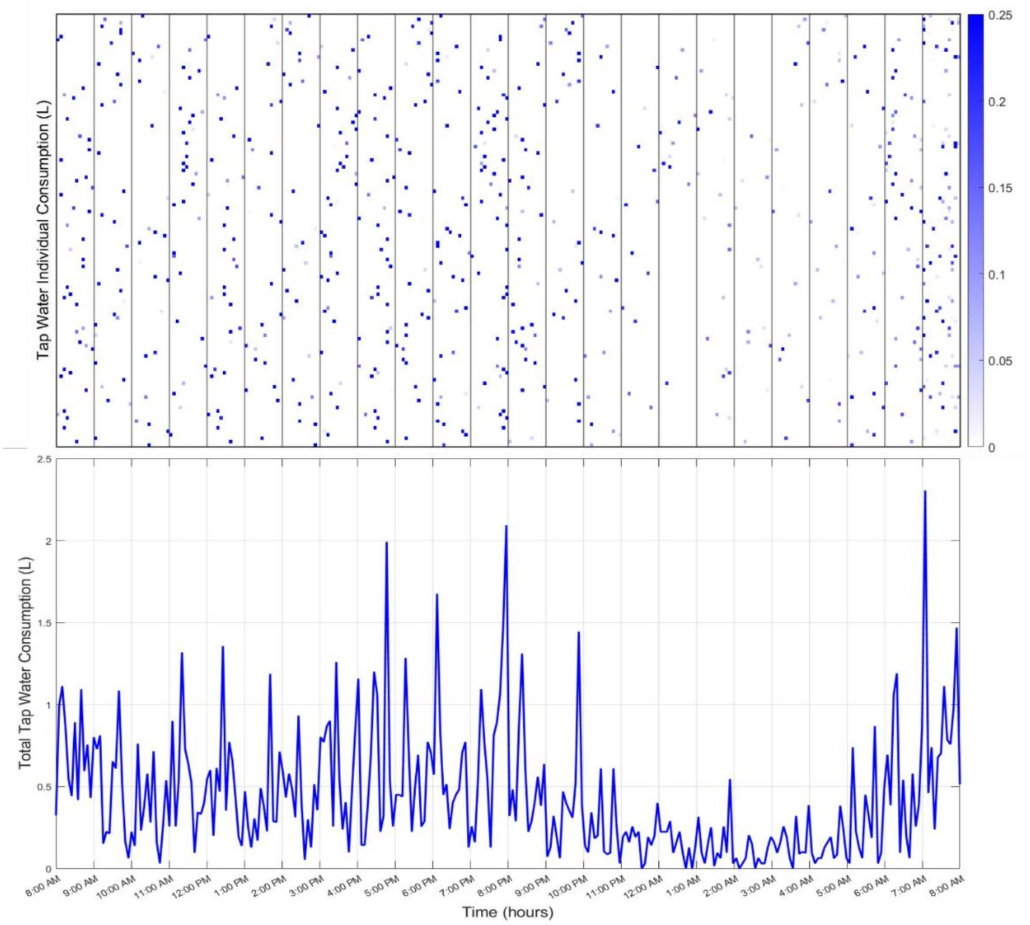


Figure 4.3 The individual daily tap water consumption and the total tap water consumption for a specific node in L-Town. Each rectangle in the heatmap represents an individual consumption event, while the color indicates the volume of water consumed.

4.3.2 Health effects assessment

Each pathogen is characterized by a unique dose-response, reflecting their individual levels of infectivity. Dose response of enterovirus (coxsackie) is commonly calculated with an exponential model utilizing a value of 0.14772 for the probability of microorganism survival r (Chigor et al., 2014). For *Campylobacter* and *Cryptosporidium*, we follow Teunis et al. (2018) and Sterk et al. (2016), using the Beta-Poisson dose-response model with the hypergeometric (1F1) function for the probability of infection from outbreak studies. The parameters α and β are 0.38 and 0.51 for *Campylobacter*, and 0.106 and 0.295 for *Cryptosporidium* respectively.

4.3.3 Risk Characterization

Integrating exposure and health effects data, we can calculate the infection risk for a specific contamination scenario. The infection probability is first calculated per individual ind, at a node v_i, at each time step k, considering a number of exposure events E_{ind} for each individual. Following WHO's approach (World Health Organization, 2016), we calculate infection probability from multiple exposure events over time as follows:

$$P_{inf}(ind) = 1 - \prod_{1}^{E_{ind}} \left(1 - P_{inf}(E, ind)\right) \tag{13}$$

where $P_{inf}(ind)$ is the infection probability of a single individual over the course of the contamination scenario given E_{ind} exposure events, and $P_{inf}(E, ind)$ is the probability of infection from a single exposure event $E \in \{1, \dots, E_{ind}\}$, derived from the dose-response function of the relevant pathogen. Note that exposure events vary per individual. The number of expected infections per node v_i is then given by the sum of probabilities of infection for each individual at the node:

$$N_{inf,i} = \sum_{ind=1}^{Pop_{(i)}} P_{inf}(ind) \qquad 14$$

The infection risk for the total population in the network, given a contamination scenario, is the ratio of the total expected infected population to the total population $Pop = \sum_{i=1}^{n_v} Pop_{(i)}$, as follows:

$$R = \frac{1}{Pop} \sum_{i=1}^{n_v} N_{inf,i} \qquad 15$$

In addition, we also evaluate the infection risk for the population at the downstream nodes of the contamination source. Let $Pop_{exp} < Pop$ be the number of individuals at contaminated nodes; then, the infection risk of the exposed population R_{exp} is:

$$R_{exp} = \frac{1}{Pop_{exp}} \sum_{i=1}^{n_v} N_{inf,i} \qquad 16$$

4.4 Contamination scenarios

Both chlorinated and non-chlorinated networks began the contamination at 08:00 AM with a temperature of 12 $°C$. A constant SRA injection of 1.85 mgCl-equiv/L was introduced from the entry points. The chlorinated network also received a constant chlorine injection of 0.5 mg/L from the entry points.

Initial pathogen concentrations were selected as mean values based on typical concentrations in raw wastewater assuming all culture-based data represent infectious pathogens (Pitkänen and Hänninen, 2017; Betancourt and Shulman, 2016; Betancourt, 2019). The contamination duration was set at 8 hours, with a wastewater injection rate of 100 L/h to represent large contamination, while dilution was calculated from the water flow at the node where the intrusion of wastewater was modeled.

4.4.1 Contamination location

The study expects contamination location to significantly impact health risks, as network nodes have varying hydraulics and affect different population levels over time (Pelekanos et al., 2021). We divided the network into three zones, choosing three contamination locations for the main scenario (Figure 4.4) based on their downstream population. The first location could affect about 50% of the population (Loc-L), the second around 30% (Loc-M), and the third a smaller segment at 10% (Loc-S).

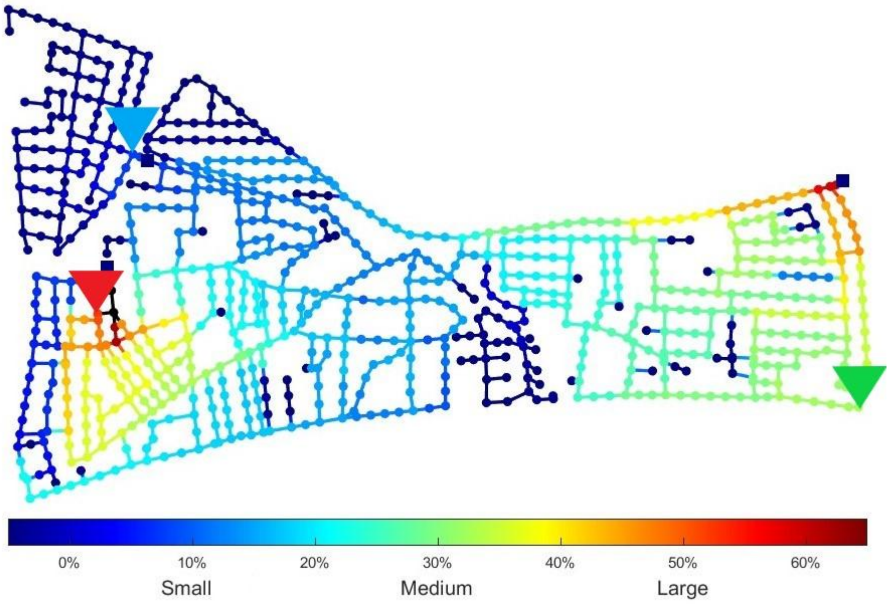


Figure 4. 4 The three selected contamination locations for the main scenario with small (Loc-S), medium (Loc-M) and large (Loc-L) potential impact on the population.

4.4.2 Hydraulic uncertainty

Most modeling studies simplify water network hydraulics by using nominal demands, overlooking the dynamic, uncertain nature of water demands. To address this, we vary the average (base) demand of each node randomly between ±10% of the nominal value and then generate a stochastically determined water demand for every node using STREaM. This procedure is reiterated 100 times generating 100 unique demand profiles for each node. The goal is to ascertain whether hydraulic uncertainty influences the model outcome for each of the three contamination locations using the pathogen *Cryptosporidium*.

4.4.3 Quality dynamics variability and uncertainty

Modeling quality dynamics requires considering input parameter variability or uncertainty. For both enterovirus and *Campylobacter*, various inactivation rates and concentrations exist in wastewater under different conditions. To understand the impact of these uncertainties on the model outcome, we conducted a nominal range sensitivity analysis on 1) Pathogen inactivation rate (using only lower rates as higher rates eliminate all pathogens), 2) Initial contaminant concentration (considering only the highest reported concentration), and 3) Contamination duration (ranging from 2 to 24 hours). Table 4.2 summarizes the quality parameters and initial conditions for the contamination scenario and sensitivity analysis.

Table 4. 2 Quality parameters and initial conditions incorporated into the benchmark model.

Parameter	**Units**	**Contamination scenario**	**Sensitivity analysis**	**Notes**	**Reference**
$C_{Cl}\ (u^{(1)})$	mg/L	0.5	0.5	Initial chlorine concentration	-
FRA ($\phi^{(2)}$)	mgCl-equiv	$0.4TOC$	$0.4TOC$	Fast-reducing agent	Fisher et al., 2017a
SRA ($\phi^{(3)}$)	mgCl-equiv	$0.6CRA$	$0.6CRA$	Slow chlorine-reducing agent	Fisher et al., 2017a
$C_{SRA}\ (u^{(2)})$	mgCl-equiv/L	$0.6CRA$	$0.6CRA$	Slow chlorine-reducing agent from reservoir	Fisher et al., 2017a
k_{FRA}	L/mgh	0.28	0.28	decay rate coefficients for fast reactions	Monteiro et al., 2014
k_{SRA}	L/mgh	0.007	0.007	decay rate coefficients for slow reactions	Monteiro et al., 2014
T	Celsius	12	5 (Correlated with inactivation rate)	Temperature	-
$C_{P_1}(\phi^{(1)})$	PFU/L	$1.39e+06$	$2.08e+07$	Enterovirus initial concentration	Betancourt and Shulman, 2016
$C_{P_2}(\phi^{(1)})$	CFU/L	$9.02e+06$	$6.2e+07$	*Campylobacter* initial concentration	Pitkänen and Hänninen, 2017
$C_{P_3}(\phi^{(1)})$	oocysts/L	$3.54e+07$	$5.4e+08$	*Cryptosporidium* initial concentration	Betancourt, 2019
TOC	mg/L	140	250	TOC concentration in wastewater	Metcalf et al., 1991; Henze et al., 2002
k_{p1}	L/mg h	92.3	-/19.4	Enterovirus inactivation rate	Rachmadi et al., 2020
k_{p2}	L/mg h	265.8	-/157	*Campylobacter* inactivation rate	Rachmadi et al., 2020
k_{p3}	L/mg h	0	0	*Cryptosporidium* inactivation rate	-
Duration	Hours	8	$2/24$	Contamination duration	-
A	dm/h	$0.01-1$	$0.01-1$	Amplification factor	Monteiro et al., 2020
B	L/mg	14	14	Rate coefficient	Monteiro et al., 2020
Km	Ft/h	$1.5826e-04 * RE^{(0.58/D)}$	$1.5826e-04 * RE^{(0.58/D)}$	Mass transfer coefficient	Shang et al., 2008b

4.5 Results and discussion

4.5.1 Chlorinated network

Figure 4.5 shows the average chlorine distribution in the L-Town network. A chlorine concentration of 0.5 mg/L is added at the two entry points, to maintain adequate residual chlorine. While this goal is mostly achieved, a concerning area arises in the northwest where chlorine levels drop critically. This area is characterized by a flow-controlling pump that fills a tank during nighttime hours, which is then used for the morning water demand, affecting the chlorine levels.

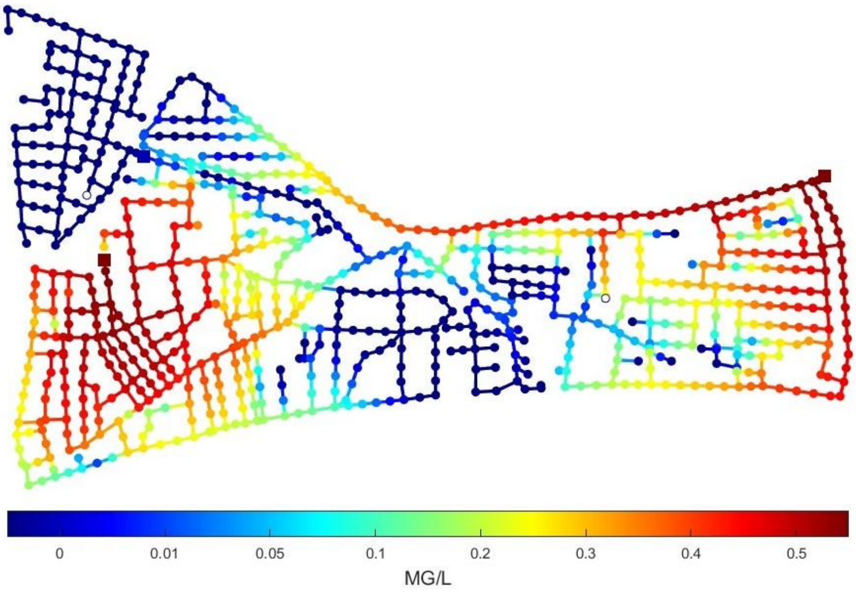


Figure 4. 5 The chlorine residual in the L-Town network under normal operation.

Table 4.3 evaluates contamination scenarios in the chlorinated network. At the Loc-L and Loc-M locations, near the chlorinated entry points of the DWDN, chlorine residual hinders pathogen spread, resulting in fewer infections and lower infection risk for both *Campylobacter* and enterovirus.

Table 4. 3 The results for the three pathogens in the chlorinated network. N_{inf} is the total infections, R is the infection risk for the total population, and R_{exp} is the infection risk of the downstream affected population.

Source Location	**Loc-L**			**Loc-M**			**Loc-S**		
Pathogens	N_{inf}	R	R_{exp}	N_{inf}	R	R_{exp}	N_{inf}	R	R_{exp}
Campylobacter	318	0.95%	2.1%	83	0.25%	0.78%	3724	11.2%	100%
Enterovirus	1158	3.5%	7.8%	2793	8.4%	26.6%	3724	11.2%	100%
Cryptosporidium	15002	45.1%	97.2%	10268	30.9%	97.9%	3705	11.1%	99.6%

Despite the higher *Campylobacter* concentration in wastewater compared to enterovirus (Table 4.2), the infection risk from enterovirus is higher in both Loc-L and Loc-M. This is due to *Campylobacter's* higher inactivation rate, reducing its concentration (upon chlorine reaction) more than enterovirus. This occurs even though the dose response relationship suggests that *Campylobacter* is more infectious than enterovirus when both are present at the same concentration (Figure 4.6).

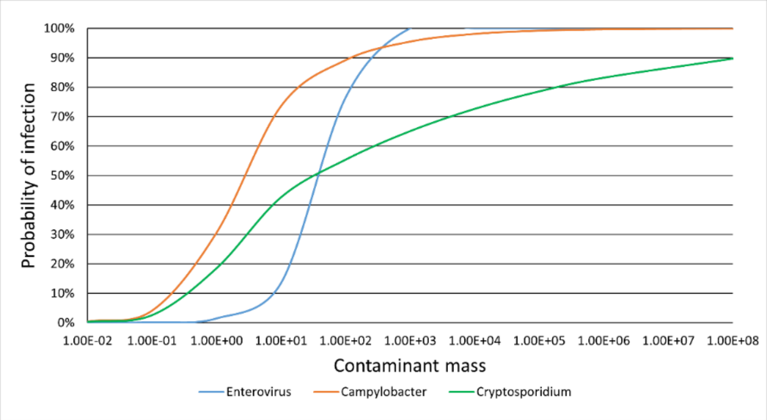


Figure 4. 6 The dose-response of the three reference pathogens.

At the Loc-S location, low chlorine levels are due to both bulk and wall decay. Extended travel time causes more chlorine decay before chlorine reaches Loc-S. During wastewater intrusion, less dilution leads to higher FRA levels entering the network causing rapid chlorine decay. Simultaneously, less diluted pathogen concentrations increase the number of initial and surviving pathogens, raising exposure to all pathogens downstream. This is causing a higher number of infections and infection risk. Figures S4.2-S4.7 in the supplementary material show a 24-hour chlorine residual profile along with *Campylobacter* and enterovirus concentration for a node downstream of each of the three contamination locations. Among the three pathogens, *Cryptosporidium* poses the greatest risk, showing the highest infection risk in Loc-L and Loc-M locations. This elevated risk profile can be attributed to *Cryptosporidium's* resistance to chlorine disinfection, showing chlorine provides no protection against chlorine-resistant pathogens.

Figure 4.7 shows the infection risk over 24 hours since contamination. For pathogens *Campylobacter* and enterovirus, the infection risk primarily emerges from the Loc-S location. The identical Loc-S infection risk profiles for these two pathogens is due to their high concentrations, infecting the entire population.

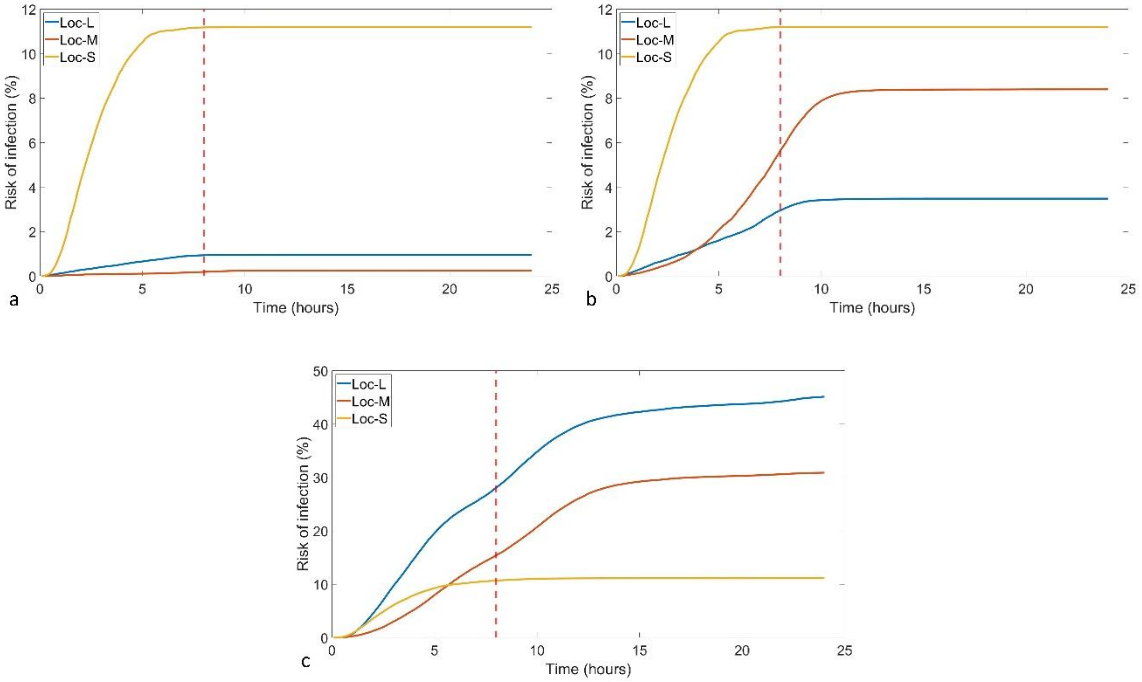


Figure 4. 7 The distribution of the infection risk in the chlorinated network for pathogens Campylobacter (a), enterovirus (b) and Cryptosporidium (c). The dashed red line indicates the end of the contamination.

The infection risk for *Cryptosporidium* differs significantly from the other two pathogens. In the Loc-L location, infection risk rapidly escalates to 25% within the first 8 hours of contamination, leveling off just above 40% by early next morning (04:00 AM). The rise after 04:00 AM is due to increased water demand as people start their (next) day, causing residual pathogens to spread and infect more individuals. The Loc-M and Loc-S locations show somewhat similar infection risks (at 15% and 10% respectively) in the first 8 hours but follow different trends. The differing risk profiles are due to variations in dilution and pathogen spread to downstream nodes. The Loc-M location, having higher dilution in certain areas and a longer path for pathogens to reach downstream nodes, experiences a gradual risk increase. Conversely, the Loc-S location has less dilution and quicker pathogen reach to downstream nodes, resulting in a more immediate surge in the infection risk. The dilution factor also influences the infection risk profile of enterovirus for Loc-L. This is due to higher contaminant dilution, as locations near the reservoir serve more downstream nodes and thus have increased flow. Consequently, the contaminant dose is reduced, leading to a lower infection risk compared to the Loc-M contamination location.

4.5.2 Non-Chlorinated network

Table 4.4 shows the contamination scenario without chlorine in the network. As expected, *Campylobacter* and enterovirus present a different profile than in the chlorinated scenario, while *Cryptosporidium*'s results remain the same.

Table 4. 4 The results for the three pathogens in the non-chlorinated network. N_{inf} is the total infections, R is the infection risk for the total population, R_{exp} is the infection risk of the downstream affected population.

Source Location	**Loc-L**			**Loc-M**			**Loc-S**		
Pathogens	N_{inf}	R	R_{exp}	N_{inf}	R	R_{exp}	N_{inf}	R	R_{exp}
Campylobacter	15439	46.4%	100%	10467	31.45%	99.7%	3724	11.2%	100%
Enterovirus	15041	45.2%	97.4%	10304	31%	98.2%	3724	11.2%	100%
Cryptosporidium	15002	45.1%	97.2%	10268	30.86%	97.9%	3705	11.13%	99.6%

Campylobacter shows slightly more infections and higher infection risk in the Loc-L and Loc-M locations compared to enterovirus, as seen in Figure 4.8. Although *Campylobacter* and enterovirus seem similar initially, they diverge after 7 hours. This behavior can be attributed to the initial concentration and dilution. *Campylobacter* has an initial concentration nearly 10 times higher than that of enterovirus, resulting in higher doses in the dose-response (Figure 4.6). Regarding dilution, it takes approximately 7 hours for the contaminated plume to mix with clean water originating from the network's east side. After this interplay, dilution occurs, which reduces the dose. Referring to Figure 4.6, it is evident that when the dose shifts to the left, the probability of infection from *Campylobacter* exceeds that of enterovirus at the same dose.

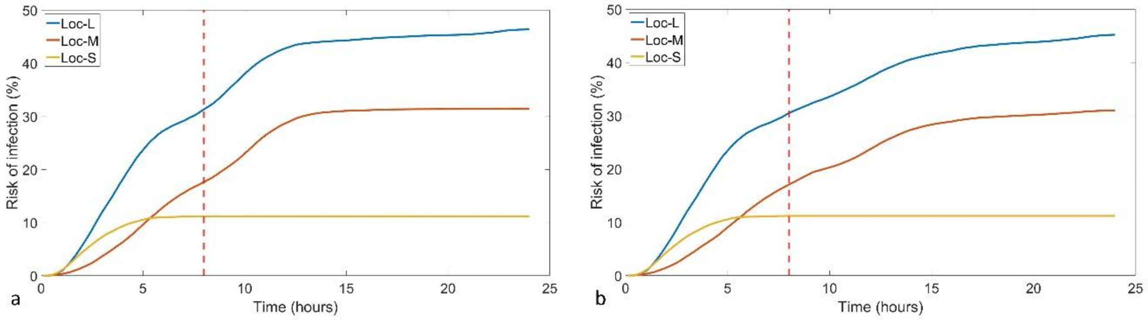


Figure 4. 8 The distribution of pathogens Campylobacter (a), enterovirus (b) depicting the infection risk in the non-chlorinated network. The dashed red line indicates the end of the contamination.

4.5.3 Hydraulic uncertainty

Figure 4.9 presents the temporal progression of the infection risk over a 24-hour period for the pathogen *Cryptosporidium*, as analyzed across 100 hydraulic scenarios. The influence of hydraulic uncertainty on the estimated infection risk is highlighted in all three locations. Specifically, both the Loc-L and Loc-M locations show approximately 3% variability, whereas the Loc-S location's variability remains under 1%. This 3% variability represents a difference of about 1000 infections. Had we incorporated a larger degree of population uncertainty before calculating water demand we expect to have seen more variability.

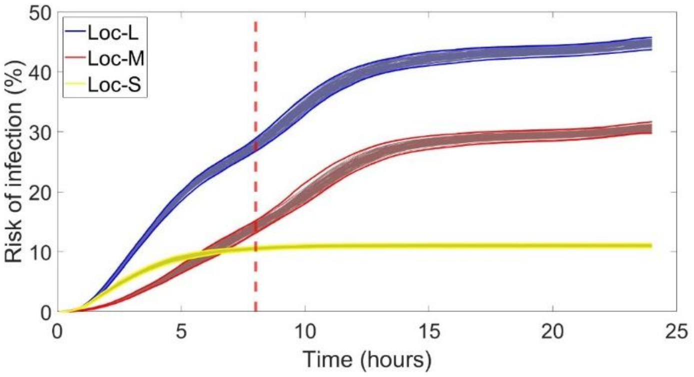


Figure 4. 9 Progression of the infection risk through a 24-hour period for the 100 hydraulic profiles. The dashed red line indicates the end of the 8-hour contamination.

4.5.4 Nominal range sensitivity analysis

Figure 4.10 shows the infection risk profile of *Campylobacter* and enterovirus in the Loc-L location for the 3 parameters of the sensitivity analysis. Figure S4.1 in the supplementary material presents the sensitivity analysis outcomes for all three locations.

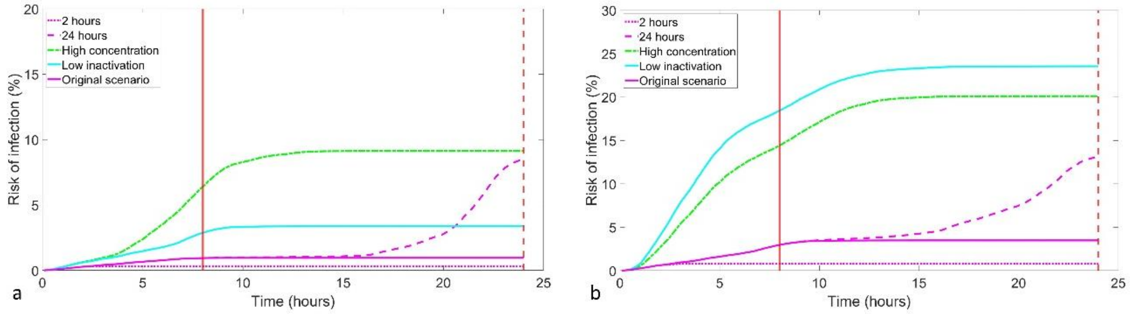


Figure 4. 10 The sensitivity analysis infection risk of Campylobacter (a) and enterovirus (b) in the Loc-L location. The solid and dashed red lines indicate the end of the 8-hour and 24-hour contamination respectively. We see that Campylobacter's infection risk is more sensitive to the initial high concentration, while enterovirus, to the low inactivation rate.

The reason the two pathogens have different impactful parameters is due to their inherent characteristics in the initial contamination scenario. For example, at the Loc-L location, despite high initial *Campylobacter* concentration, chlorine's quick inactivation leads to a lower dose, thus lower infection probability as shown in Figure 4.6. However, when the initial concentration increases even more, so does the dose and infection probability, putting a larger population at risk and leading to a significant rise in infection risk. For enterovirus, although subjected to chlorine inactivation in the original scenario, it caused more infections than *Campylobacter* due to its slower inactivation rate. A further decrease in this rate allows more contaminants into the network, substantially increasing the infection risk. Recognizing the potential variability of these parameters is crucial. Pathogen concentrations in wastewater vary, contamination events can last varying durations, and factors like temperature affect pathogen inactivation rates.

4.5.5 Role of chlorination

The results of the contamination scenarios emphasize chlorine's role in controlling waterborne pathogens in DWDN. With adequate chlorine residual, the spread of *Campylobacter* and enterovirus is greatly reduced, irrespective of contamination location as the infection risk for downstream population is only 0.78%-26.6%. However, areas far from chlorination points may lack sufficient residual chlorine. In wastewater contamination, pathogens, especially viruses with low inactivation rates, pose a notable risk. Chlorine-resistant pathogens like *Cryptosporidium also* present a threat. This demonstrates the need for a prompt response to such events, as within 5-10 hours post-contamination, 10%-35% of the entire population could be infected.

Exploring a non-chlorinated network offers a contrasting image, where the infection risk escalates, specifically with pathogens like *Campylobacter* that have typically high concentrations in wastewater. A contamination event in a non-chlorinated network could infect 97%-100% of the downstream affected population.

4.5.6 Role of contamination location

In both chlorinated and non-chlorinated networks, different contamination locations exhibit unique risk profiles. Contaminations in larger zones, especially in non-chlorinated networks, bear more infection risk over time due to more contaminated downstream nodes. Conversely, contaminations in smaller zones have lower infection risk but are prone to immediate risk surges due to rapid pathogen spread to downstream nodes. Dilution also matters; contamination near the reservoir may lead to more dilution, reducing the contaminant mass per node, while contamination at the network's periphery may result in less dilution and higher contaminant mass. This demonstrates how the network's structural characteristics influence the infection risk.

4.5.7 Limitations and recommendations

Our risk estimation relies on specific dose-response models. Using different dose-response models could alter results, showing that the computed risk is tied to the chosen dose-response model. A limitation is the study's focus on individual pathogen infection risk. Realistically, wastewater carries multiple pathogens at different concentrations. Future work should explore cumulative infection risk, considering the dose response of all pathogens together for a holistic infection risk assessment, especially in networks with low to no chlorine residual.

One form of limitation is the assumption that exposure to pathogens occurs only through kitchen tap water ingestion since exposure can also happen via showering, brushing teeth. Although our proposed methodology for the distribution of daily tap water consumption provides adequate variability of drinking water consumption over time as evidenced in Figure 4.3, exploring variability of individual consumption volumes of tap water (e.g., 0, 0.2, 0.5, 2, 3 or 4 L/p/d) could capture even more variability in individual water consumption behavior.

Another limitation is the absence of real data for CRA concentrations in wastewater. Our analysis uses approximations made in the literature that consider TOC concentration as representative of CRA, assuming chlorine mainly interacts with inorganic substances (and some organic) during the fast phase, and mainly organics in the slow phase.

The calibration of the BeWaRE testbed was carried out using network-specific data that enhanced the accuracy of a model. Several components of the BeWaRE testbed are transferable and can be applied to different networks (with different types of water systems), e.g., the water quality model component, or the QMRA component. However, if water utilities want to use those components in their network, they would first need to calibrate the parameters to get realistic results.

4.5.8 Operational changes in water management and policy

The proposed model can potentially influence operational changes in water management since it emphasizes the necessity of advanced modeling tools to effectively mitigate pathogen contamination events in the DWDN. Not many people are exposed to incidents but when they are affected by an incident, their exposure to pathogens (or infection risk) is high and there is a small window of opportunity for meaningful interventions. We believe that our tool can also influence policy. Once again, the results of our contamination scenario highlight the need to integrate Water Security Plans into existing Water Safety Plans and to develop Standard Operating Procedures for contamination emergency responses. Finally, using such computational models to estimate (with high resolution) health impact and adopting the use of such technologies for decision support can optimize response strategies and improve system resilience.

4.5.9 Application to real case studies

BeWaRE has been applied to real case studies in the context of the EU-funded PathoCERT (Pathogen Contamination Emergency Response Technologies) project. The aim of the project was to enhance the coordination capabilities of first responders during pathogen contamination emergencies. BeWaRE was integrated into a decision-support tool named PathoINVEST (Paraskevopoulos et al., 2022) and was applied in three European case studies, each featuring distinct characteristics. In Spain and Cyprus, it was used to assist the response to earthquakes that led to sewage infiltration into the DWDN. In the Netherlands, it was employed to investigate suspected intentional contamination following customer complaints. In each case study, emergency response teams comprising individuals from all relevant sectors (water utilities, civil protection, and health care). These teams incorporated their own network data into BeWaRE, having an accurate representation of the contaminant transport, as well as health impact analysis.

4.6 Conclusion

This study introduced BeWaRE, an open-access testbed, featuring an integrated hydraulic and water quality model that can be seamlessly incorporated into any DWDN model once properly calibrated. It brings together key factors -chlorine decay (bulk and wall), fast and slow reactions, CRA degradation, and stochastic water use patterns- into a single model within a QMRA framework. By integrating these elements, it provides a more realistic and reliable way to assess infection risks during contamination events, addressing limitations of previous models that relied on simplified assumptions. Its applicability was demonstrated by analyzing the health impacts of wastewater contamination in a DWDN and evaluating the role of chlorination in mitigating risks from different enteric pathogens. Key findings include:

- In non-chlorinated DWDN, the modeled wastewater contamination events led to 11%-46% infection risk in the total population, depending on the contamination location, but irrespective of the selected pathogen (due to the high pathogen concentration).
- In chlorinated DWDN, the same scenarios resulted in lower infection risk for the pathogens that are susceptible to chlorine; 0.78%-2.1% for *Campylobacter* and 7.8%-26.6% for enterovirus. Enterovirus infection risk is higher, despite the concentrations in the contamination source being lower, due to the lower susceptibility to chlorine than *Campylobacter.*
- In chlorinated DWDN, the modeled contamination scenarios yielded infections as a result of Cryptosporidium, due to its high chlorine resistance. Contamination location plays a significant role in terms of impact, due to the size of the affected population, but also due to the level of dilution of the contamination in the DWDN.
- The response window after a contamination event to reduce the health impact is short; in these scenarios 5-10 hours post-contamination.
- *Campylobacter's* infection risk is more sensitive to the initial concentration in the contamination source whereas enterovirus infection risk to inactivation rate.

This testbed can serve as a baseline for future studies, potentially exploring different inactivation kinetics, alternative pathogens, varied water consumption patterns, or calibrating the model for real-world drinking water systems. This study illuminates the profound health implications of a large wastewater contamination in drinking water networks. While chlorination plays an essential defensive role, a comprehensive understanding of pathogen behavior is crucial for enhancing protection against potential outbreaks and ensuring a safer water supply.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 883484 (PathoCERT), the European Research Council (ERC), under the ERC Synergy grant agreement No. 951424 (Water Futures), and was supported by the European Union’s Horizon 2020 Teaming programme under grant agreement No. 739551 (KIOS CoE), and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

Supplementary material

4.2.6 Modeling reactions

The L-Town benchmark network utilized in this study is based on a real District Metered Area (DMA) network located in Cyprus. In the real network, six chlorine sensors are installed: one at the inlet (since the network has only one inlet) and five within the network itself.

To calibrate and validate our water-quality model, we first utilized the hydraulic model of the real network, which was recalibrated using the available flow and pressure data. The water-quality model parameters were then calibrated to minimize prediction error using one week of chlorine concentration data recorded at five-minute intervals from the six sensors. Specifically, we adjusted the chlorine decay parameters, including the reactions with Slow Reacting Agents (SRA), the wall reaction amplification factor, A, and the rate coefficient B. This process ensured that the water-quality model was tailored to the specifics of the real network used, and in extension L-Town. Figure S1 shows the real and estimated chlorine levels at the six chlorine sensor locations. The model performed well in predicting both chlorine concentration levels and complete chlorine depletion at the monitoring nodes, which is crucial for the contamination impact study. Some mismatches in chlorine concentration, particularly on the first day, can be attributed to the initial uncertainty in chlorine concentration across the network and the inherent variability in hydraulic modeling.

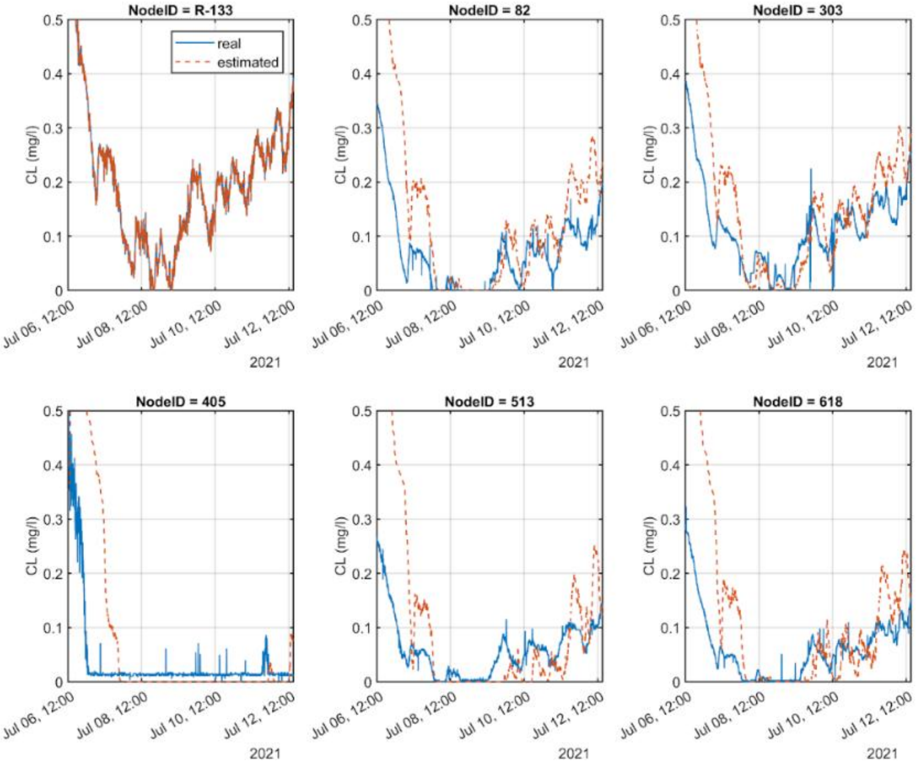


Figure S4. 1 Comparison of real and simulated chlorine concentration in the original case study network, after the proposed quality model calibration. Node R-133 is the DMA inlet node, while the rest of the nodes are locations with installed chlorine sensors.

4.5.1 Chlorinated network

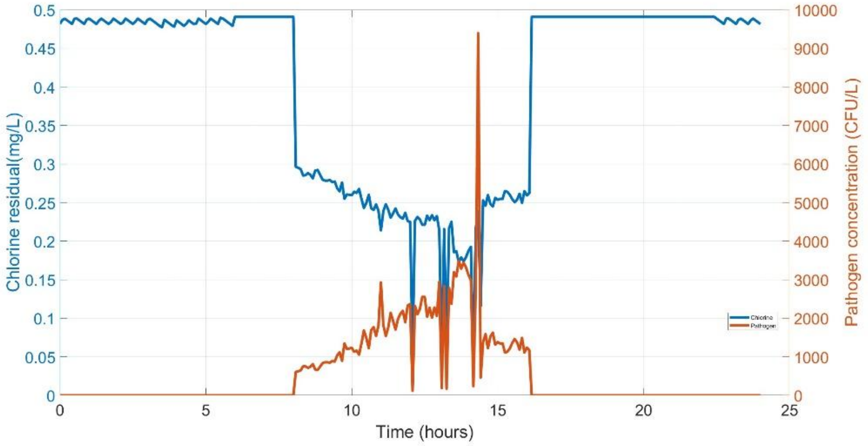


Figure S4. 2 The chlorine residual profile along with Campylobacter concentration for a node downstream of the Loc-L location.

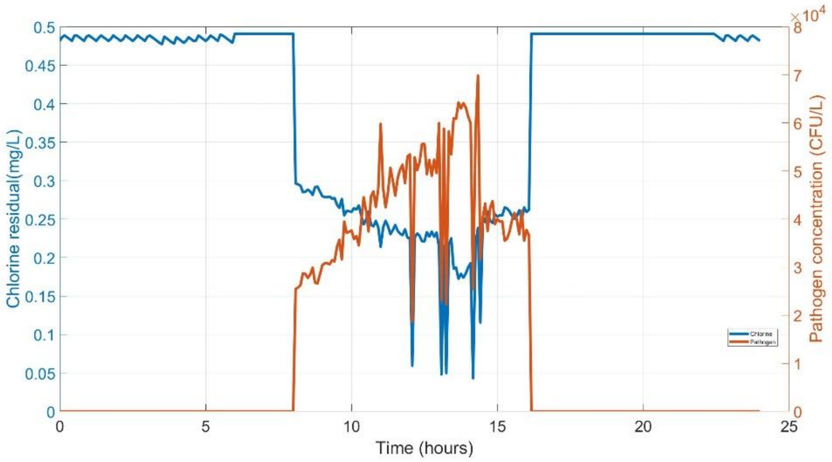


Figure S4. 3 The chlorine residual profile along with enterovirus concentration for a node downstream of the Loc-L location.

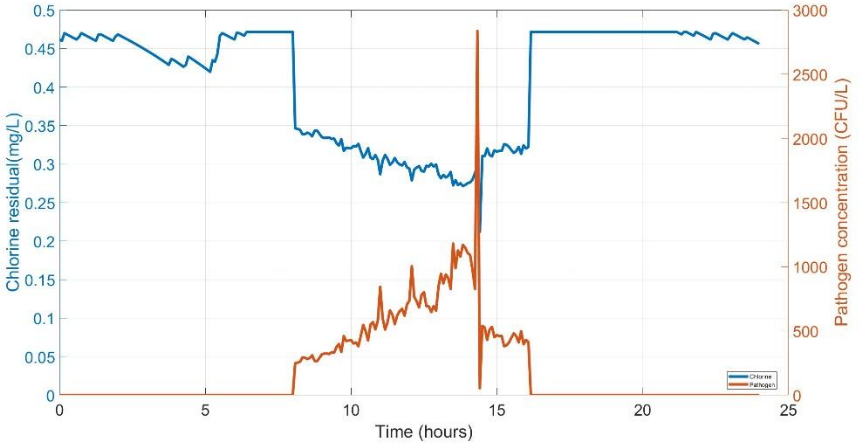


Figure S4. 4 The chlorine residual profile along with Campylobacter concentration for a node downstream of the Loc-M location.

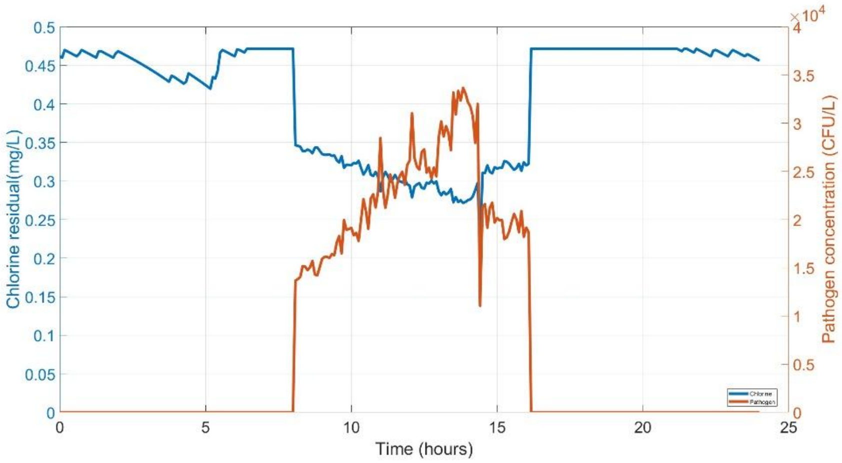


Figure S4. 5 The chlorine residual profile along with enterovirus concentration for a node downstream of the Loc-M location.

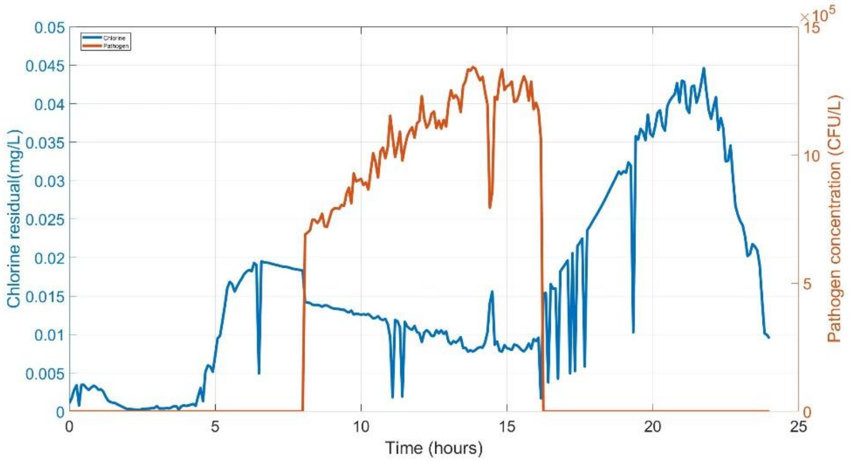


Figure S4. 6 The chlorine residual profile along with Campylobacter concentration for a node downstream of the Loc-S location.

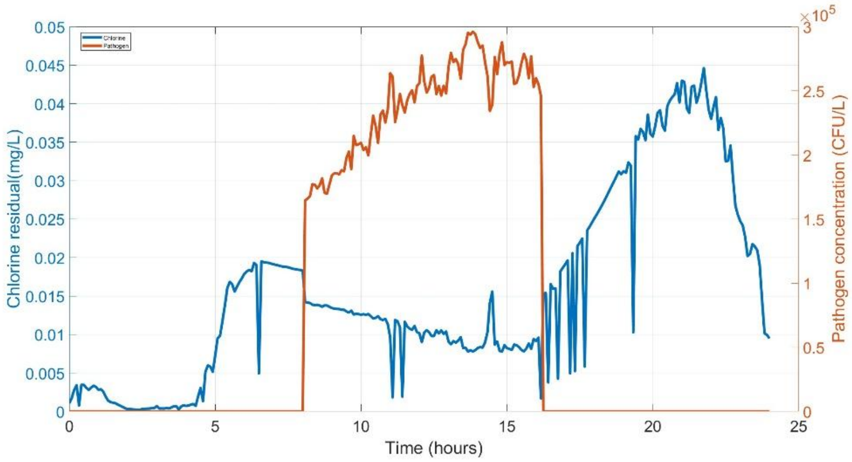


Figure S4. 7 The chlorine residual profile along with enterovirus concentration for a node downstream of the Loc-S location.

4.5.4 Nominal range sensitivity analysis

Table S4. 1 Results of the nominal range sensitivity analysis. The values represent the percentage change of infection risk relative to the initial contamination scenario.

	Campylobacter			**Enterovirus**		
	Loc-L	**Loc-M**	**Loc-S**	**Loc-L**	**Loc-M**	**Loc-S**
Duration: 2 hours	−67.7%	−76%	−6%	−77.3%	−85%	−5.9%
Duration: 24 hours	+790%	+2252%	-	+277%	+31%	-
High concentration	+861%	+3715%	-	+477%	+95%	-
Low inactivation rate	+255%	+3220%	-	+576%	+137%	-

References

Abhijith, G. R., & Ostfeld, A. (2021). Modeling the response of nonchlorinated, chlorinated, and chloraminated water distribution systems toward arsenic contamination. *Journal of Environmental Engineering*, *147*(10), 04021045.

Betancourt, W. (2019). Cryptosporidium spp.

Betancourt, W. Q., & Shulman, L. M. (2016). Polioviruses and other Enteroviruses. *Global Water Pathogen Project; Rose, JB, Jiménez-Cisneros, B., Eds*.

Betanzo, E. W., Hofmann, R., Hu, Z., Baribeau, H., & Alam, Z. (2008). Modeling the impact of microbial intrusion on secondary disinfection in a drinking water distribution system. *Journal of Environmental Engineering*, *134*(4), 231-237.

Blokker, M., Smeets, P., & Medema, G. (2018). Quantitative microbial risk assessment of repairs of the drinking water distribution system. *Microbial Risk Analysis*, *8*, 22-31.

Boulos, P. F., Lansey, K. E., & Karney, B. W. (2006). Comprehensive water distribution systems analysis handbook for engineers and planners. American Water Works Association.

Chigor, V. N., Sibanda, T., & Okoh, A. I. (2014). Assessment of the risks for human health of adenoviruses, hepatitis A virus, rotaviruses and enteroviruses in the Buffalo River and three source water dams in the Eastern Cape. *Food and Environmental Virology*, *6*, 87-98.

Clark, R. M., Yang, Y. J., Impellitteri, C. A., Haught, R. C., Schupp, D. A., Panguluri, S., & Krishnan, E. R. (2010). Chlorine fate and transport in distribution systems: Experimental and modeling studies. *Journal-American Water Works Association*, *102*(5), 144-155.

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D. E., & Abdallah, A. M. (2018). Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. *Environmental Modeling & Software*, *102*, 199-212.

Craun, G. F., & Calderon, R. L. (2001). Waterborne disease outbreaks caused by distribution system deficiencies. *Journal-American Water Works Association*, *93*(9), 64-75.

Douterelo, I., Husband, S., Loza, V., & Boxall, J. (2016). Dynamics of biofilm regrowth in drinking water distribution systems. *Applied and environmental microbiology*, *82*(14), 4155-4168.

Eliades, D. G., Kyriakou, M., Vrachimis, S., & Polycarpou, M. M. (2016, November). EPANET-MATLAB toolkit: An open-source software for interfacing EPANET with MATLAB. In *Proc. 14th international conference on computing and control for the water industry (ccwi)* (Vol. 8).

Eliades, D. G., Vrachimis, S. G., Moghaddam, A., Tzortzis, I., & Polycarpou, M. M. (2023). Contamination event diagnosis in drinking water networks: A review. *Annual Reviews in Control*, 55, 420-441.

Fisher, I., Kastl, G., & Sathasivan, A. (2017a). A comprehensive bulk chlorine decay model for simulating residuals in water distribution systems. *Urban Water Journal*, *14*(4), 361-368.

Fisher, I., Kastl, G., & Sathasivan, A. (2017b). New model of chlorine-wall reaction for simulating chlorine concentration in drinking water distribution systems. *Water research*, *125*, 427-437.

Frankel, M., Katz, L. E., Kinney, K., Werth, C. J., Zigler, C., & Sela, L. (2023). A framework for assessing uncertainty of drinking water quality in distribution networks with application to monochloramine decay. *Journal of Cleaner Production*, *407*, 137056.

Giammanco, G. M., Bonura, F., Urone, N., Purpari, G., Cuccia, M., Pepe, A., ... & De Grazia, S. (2018). Waterborne Norovirus outbreak at a seaside resort likely originating from municipal water distribution system failure. Epidemiology & Infection, 146(7), 879-887.

Henze, M., Harremoes, P., Jansen, J. L., & Arvin, E. (2002). *Wastewater treatment: biological and chemical process*. Springer-Verlag.

Hrudey, S. E., & Hrudey, E. J. (2004). *Safe drinking water*. IWA publishing.

Hrudey, S. E., & Hrudey, E. J. (2019). Common themes contributing to recent drinking water disease outbreaks in affluent nations. *Water supply*, *19*(6), 1767-1777.

Kuhn, K. G., Falkenhorst, G., Emborg, H. D., Ceper, T., Torpdahl, M., Krogfelt, K. A., ... & Mølbak, K. (2017). Epidemiological and serological investigation of a waterborne Campylobacter jejuni outbreak in a Danish town. Epidemiology & Infection, 145(4), 701-709.

Laine, J., Huovinen, E., Virtanen, M. J., Snellman, M., Lumio, J., Ruutu, P., ... & Pirkanmaa Waterborne Outbreak Study Group. (2011). An extensive gastroenteritis outbreak after drinking-water contamination by sewage effluent, Finland. *Epidemiology & Infection*, *139*(7), 1105-1113.

LeChevallier, M. W. (1999). The case for maintaining a disinfectant residual. *Journal-American Water Works Association*, *91*(1), 86-94.

Medema, G. J., Smeets, P. W. M. H., Van Blokker, E. J. M., & Lieverloo, J. H. M. (2013). Safe distribution without a disinfectant residual. *Microbial Growth in Drinking-Water Supplies. Problems, Causes, Control and Research Needs*, 95-125.

Metcalf, L., Eddy, H. P., & Tchobanoglous, G. (1991). *Wastewater engineering: treatment, disposal, and reuse* (Vol. 4). New York: McGraw-Hill.

Monteiro, L., Carneiro, J., & Covas, D. I. (2020). Modeling chlorine wall decay in a full-scale water supply system. *Urban Water Journal*, *17*(8), 754-762.

Monteiro, L., Figueiredo, D., Dias, S., Freitas, R., Covas, D., Menaia, J., & Coelho, S. T. (2014). Modeling of chlorine decay in drinking water supply systems using EPANET MSX. *Procedia Engineering*, *70*, 1192-1200.

Odhiambo, M., Viñas, V., Sokolova, E., & Pettersson, T. J. (2023). Health risks due to intrusion into the drinking water distribution network: hydraulic modeling and quantitative microbial risk assessment. *Environmental Science: Water Research & Technology*, *9*(6), 1701-1716.

Paraskevopoulos, S., Vrachimis, S., Kyriakou, M., Pavlou, P., Kouzapas, D., Milis, G., Smeets, P., Eliades, D., Medema, G., Polycarpou, M., & Panayiotou, C. (2022). PathoINVEST: Pathogen contamination investigations during emergencies. In 2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Indsustry. Universitat Politecnica de Valencia. https://doi.org/10.4995/WDSA-CCWI2022.2022.14799

Pelekanos, N., Nikolopoulos, D., & Makropoulos, C. (2021). Simulation and vulnerability assessment of water distribution networks under deliberate contamination attacks. *Urban Water Journal*, *18*(4), 209-222.

Pitkänen, T., & Hänninen, M. L. (2017). Members of the family Campylobacteraceae: Campylobacter jejuni, Campylobacter coli. *Global Water Pathogens Project. http://www. waterpathogens. org*.

Rachmadi, A. T., Kitajima, M., Kato, T., Kato, H., Okabe, S., & Sano, D. (2020). Required chlorination doses to fulfill the credit value for disinfection of enteric viruses in water: a critical review. *Environmental Science & Technology*, *54*(4), 2068-2077.

Rossman, L. A. (2000). EPANET 2: users manual.

Shang, F., Uber, J. G., & Polycarpou, M. M. (2002). Particle backtracking algorithm for water distribution system analysis. *Journal of environmental engineering*, *128*(5), 441-450.

Shang, F., Uber, J. G., & Rossman, L. A. (2008a). Modeling reaction and transport of multiple species in water distribution systems. *Environmental science & technology*, *42*(3), 808-814.

Shang, F., Uber, J. G., Rossman, L. A., & Janke, R. (2008b). EPANET multi-species extension user's manual. *Risk Reduction Engineering Laboratory, US Environmental Protection Agency, Cincinnati, Ohio*.

Sterk, A., Schijven, J., de Roda Husman, A. M., & de Nijs, T. (2016). Effect of climate change on runoff of Campylobacter and Cryptosporidium from land to surface water. *Water research*, *95*, 90-102.

Teixeira, R., Carmi, O., Raich, J., Gattinesi, P., & Hohenblum, P. (2019). Guidance for production of a Water Security Plan in drinking water supply.

Teunis, P. F. M., Xu, M., Fleming, K. K., Yang, J., Moe, C. L., & LeChevallier, M. W. (2010). Enteric virus infection risk from intrusion of sewage into a drinking water distribution network. *Environmental science & technology*, *44*(22), 8561-8566.

Teunis, P. F., Marinović, A. B., Tribble, D. R., Porter, C. K., & Swart, A. (2018). Acute illness from Campylobacter jejuni may require high doses while infection occurs at low doses. *Epidemics*, *24*, 1-20.

Todini, E. (1987). A gradient method for the analysis of pipe networks. In *International Conference on Computer Applications for Water Supply and Distribution*.

van Lieverloo, J. H. M., Blokker, E. M., & Medema, G. (2007). Quantitative microbial risk assessment of distributed drinking water using faecal indicator incidence and concentrations. *Journal of Water and Health, 5(S1), 131-149.*

Vieira, P., Coelho, S. T., & Loureiro, D. (2004). Accounting for the influence of initial chlorine concentration, TOC, iron and temperature when modeling chlorine decay in water supply. *Journal of Water Supply: Research and Technology—AQUA*, *53*(7), 453-467.

Vrachimis, S. G., Eliades, D. G., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., ... & Polycarpou, M. M. (2022). Battle of the leakage detection and isolation methods. *Journal of Water Resources Planning and Management*, *148*(12), 04022068.

Vrachimis, S. G., Timotheou, S., Eliades, D. G., & Polycarpou, M. M. (2019). Iterative hydraulic interval state estimation for water distribution networks. *Journal of Water Resources Planning and Management*, *145*(1), 04018087.

World Health Organization. (2016). Quantitative microbial risk assessment: application for water safety management.

World Health Organization. (2017). Guidelines for drinking-water quality: first addendum to the fourth edition.

Yang, J., LeChevallier, M. W., Teunis, P. F., & Xu, M. (2011). Managing risks from virus intrusion into water distribution systems due to pressure transients. Journal of water and health, 9(2), 291-305.

Chapter 5

Addressing Drinking Water Contamination: A Case Study Comparing Traditional with Model-based Approaches

Currently under review in the Journal of Water Resources Planning and Management.
Paraskevopoulos, S., Vrachimis, S., Kyriakou, M., Eliades, D., Smeets, P., Blokker, M., Polycarpou, M., Medema, G.

Abstract

Rapid and effective decision-making is crucial during drinking water contamination events to ensure public safety. This Chapter examines a case study where a water utility, responding to customer complaints, suspected wastewater contamination in its network. We compare the traditional expert judgement approach to a model-based approach using the PathoINVEST tool. The tool performs simulations of contamination events informed by sensor measurements, identifies contamination sources using sampling results, and suggests optimal valve closures for mitigation.

Our findings show that the model-based approach significantly enhances response efficiency and accuracy. It identified the contamination source with four samples in 1.3 hours, compared to 11 samples in 3.7 hours for the traditional approach, and resulted in a lower infection risk (12% vs. 20%) at the time of source identification. Regarding valve closure, the model-based approach performed better, resulting in a 3%-point reduction in infection risk compared to the traditional approach. Modeling uncertainty is addressed by considering valve settings uncertainty; despite a 0.7% discrepancy in valve settings compared to the model, the tool accurately pinpointed the contamination vicinity 75% of the time.

These findings support the claim that integrating modeling and sensor tools into emergency response protocols for drinking water contamination events can improve early identification and mitigation, potentially safeguarding public health in urban water supply systems.

Keywords: Drinking water contamination modeling; source identification; emergency response; water utility; valve manipulation

5.1 Introduction

Ensuring the safety of drinking water is a primary concern of water utilities worldwide. Contamination events in the drinking water network (DWDN) can arise from human error, infrastructure failures, main repairs, or malicious attacks (Winston et al., 2003; Hrudey and Hrudey, 2004; Blackburn et al., 2004; Arnone and Walling, 2007; Fewtrell et al., 2011; Laine et al., 2011; Cann et al., 2013; Lendowski et al., 2015; Blokker et al., 2018) directly impacting public health and well-being (Kunz, 2024). Advancements in monitoring and remediation technologies have enhanced our ability to respond to such events (Erickson et al., 2019), yet there remains a challenge in effectively using real-time modeling tools to aid responsible authorities (Eliades et al., 2023).

Despite already established protocols and tools for emergency response, pathogen-related contamination events in the DWDN continue to affect communities. Communication gaps and the underutilization of available technologies play an important role. Rapid and effective decision-making is crucial during such events; delays or inaccuracies in addressing these situations can lead to more people affected, escalated health risks, economic losses, and prolonged recovery periods (Laursen et al., 1994; Corso et al., 2003; Ailes et al., 2013; Chyzheuskaya et al., 2017; Gude and Muire, 2021). Moreover, repeated or mismanaged events erode public trust in the safety of drinking water, which is a foundation of urban living and public health (Anadu and Harding, 2000).

Currently, when a DWDN is contaminated by wastewater, water utilities activate their crisis management procedures, to find the suspected contamination source and minimize as soon as possible the impact of contamination through mitigation measures (USEPA, 2003). Traditional decision-making during the crisis management procedure has been predominantly guided by past experience, best practice industry protocols, and expert judgment based on (limited) available information on the contamination event. Regarding the knowledge of the network characteristics, water utilities rely on solvers (e.g., EPANET, WATERGEMS, etc) to model the hydraulics and water quality dynamics. While experts using those modeling tools are consulted, their input often comes after a critical window of opportunity (sometimes after 24 hours have passed). They frequently rely on outdated network characteristics and slower, generic (not dedicated to wastewater contamination events) models, without being able to model the health impact of a contamination. Therefore, in this study, a traditional approach represents the status quo of current practices in water utilities (expert judgment, past experience, offline generic modeling), while a model-based approach uses real-time modeling tools for pathogen propagation based on stochastic water demands, health risk assessment, and support in decision-making. Hence, the hypothesis of this study is that the integration of modeling tools (model-based approach) with current practices (traditional approach) can enhance responses and reduce the negative impacts of contamination events.

Recently an analytical tool named PathoINVEST (Pathogen contamination INVESTigation decision support system) (Paraskevopoulos et al., 2022) was developed as part of the EU-funded PathoCERT (Pathogen Contamination Emergency Response Technologies) project to support decision-making during emergencies. It is built as a QGIS plugin and it uses the software EPANET-MATLAB Toolkit and a benchmark hydraulic and quality model incorporating various factors (e.g., pathogen concentrations in contamination sources, pathogen inactivation and chlorine demand kinetics, and stochastic water demands) (Eliades et al., 2016). This tool not only simulates contamination events and their health impact realistically but also suggests potential sampling locations to identify the suspected source of contamination and optimal valve closure strategies for mitigation strategies.

This Chapter addresses the challenge of rapidly finding the contamination source and minimizing the population's health risk through mitigation measures, aiming to improve traditional methods by incorporating modeling tools in the decision-making. Our objective is to systematically compare the efficacy of traditional and model-based decision-making in responding to wastewater contamination events in the DWDN when the source is unknown. We focus on understanding how decision-making can be enhanced with the integration of real-time modeling tools and sensors. For the source identification, the comparison metrics include the duration of time needed to find the source, the number of samples required, and the infection risk of the urban population at the time the source was identified.

For the mitigation measures, we focus on valve manipulation so that the contamination plume can be isolated or reduced, and the comparison metric is the risk of infection. We also address the problem of uncertainty when using modeling tools. Specifically, we assess how valve settings uncertainty in the DWDN can potentially provide misleading results for decision-making (using the model-based approach). We demonstrate a fictional contamination case study, observing how responsible authorities from a water utility in the Netherlands respond to suspected wastewater contamination in their DWDN.

The main contributions of this study are:

1. Demonstration of a realistic contamination case study, revealing the actual steps water utilities take in emergencies.
2. Demonstration of a software tool that simulates real-time contamination events in DWDN and provides decision support for water utilities during an emergency.
3. A comparison between traditional and model-based decision-making for source identification and mitigation measures.
4. Evaluation of valve settings uncertainty in the DWDN and how it influences the accuracy of decisions during emergencies.
5. Quantification of the health protection and operational efficiency benefits of using a model-based approach during emergencies.

5.2 Methodology

5.2.1 Overview of the PathoINVEST tool

While a detailed description of the PathoINVEST tool can be found in Paraskevopoulos et al. (2022), for completeness, we note here some key characteristics:

1. The tool incorporates up-to-date waterborne pathogen data from contamination sources and inactivation kinetics including for enterovirus, *Campylobacter*, and *Cryptosporidium*.
2. Stochastic water demands are being used to provide a more thorough understanding of the hydraulics and isolate the water end-use of tap water to calculate the risk of infection.
3. For understanding the potential health impact, the tool employs the principles of Quantitative Microbial Risk Assessment (QMRA).

For decision-making, the tool features various methods, including sampling location suggestion, to help identify the contamination source and optimal valve closure recommendation, to mitigate the infection risk by ensuring fewer people are infected.

5.2.2 Emergency response

In the event of a contamination emergency, the primary goal of a water utility is to rapidly execute a minimal yet effective set of actions to mitigate the incident and restore normal operations. We followed the US EPA Response Protocol Toolbox (RPTB) which includes a list of recommendations on actions following such an event: a) detection of contamination; b) source identification; and c) consequence management (USEPA, 2003; Afshar and Najafi, 2014). Contamination can be detected either through complaints, manual samplings, or water quality sensor signals. The next step involves identifying the suspected source of contamination. This typically involves determining strategic locations for sampling. Current microbiological testing protocols, such as culture or RT-PCR (Reverse transcription polymerase chain reaction), necessitate that samples be transported to a laboratory for analysis. Typically, the time-to-result is approximately 4 hours (RT-PCR) to 18-24 hours (culture). The positive results from one or multiple samples provide a preliminary indication of the potential origin of the contamination. At the same time, water utilities commonly issue a ``boil water" advisory in the potentially affected area as an initial step to mitigate the health impact. Modeling the contamination (once the contamination source has been identified) provides insights into how the contamination propagates over time. As a result, authorities can issue a boil water advisory to specific areas within the network and strategically close certain valves. This action effectively isolates the contaminated area, preventing further spread (consequence management). Besides valve closure, integral parts of the consequence management step are flushing and chlorination (Poulin et al., 2008), but this Chapter focuses on the initial response. Following these steps in an emergency event enables a rapid and efficient response to safeguard public health and restore the integrity of the water supply system.

5.2.3 Case study

The case study was simulated using L-Town's DWDN as the water utility's network data was restricted due to sensitive information. L-Town is a modified network from the BattLeDIM (Battle of the Leakage Detection and Isolation Methods) competition (Vrachimis et al., 2022). L-Town is a benchmark network comprising 782 junctions and 905 pipe segments, providing water to an estimated population of 30,000. For this case study, we modeled the waterborne pathogen enterovirus, a common pathogen found in wastewater in high concentrations, for a continuous contamination of 24 hours starting at 8:00 a.m. The water supplied in the network was unchlorinated since this is a common practice in the Netherlands. The demands of the network were generated using the STochastic Residential water End-use Model (STREaM) tool (Cominola et al., 2018). This tool simulates household water end-uses, each having distinct consumption patterns and probability distributions for water use volume, use duration, daily frequency, and time of use during the day. This gives a realistic residential demand profile for the L-Town network.

An emergency tabletop exercise was conducted in which a response team dealt with a contamination scenario provided by a supervisor. The supervisor, responsible for running the simulations, was the only one aware of the actual source of contamination. The supervisor provided feedback (e.g., sampling results, visualization of contamination propagation, health impact) on the proposed activities by the response team (e.g., sampling locations, valve manipulation) while recording the time that would be required for any action. The response team was provided with data on water age and daily average flow directions in the network (Figure S5.1 in the supplementary material). The scenario consisted of multiple customer complaints (Figure 5.1) at 9:30 a.m., which led to a response from the team. The response team included an incident commander, a communications manager, advisors specializing in water quality and crisis management, and a modeler. The incident commander implemented a structured decision-making approach, consisting of the following steps: observe, assess, and decide. After reviewing the provided information on water age and flow direction, the response team decided that the contamination likely began around 8:00 a.m. in the network's eastern part. Additionally, two subgroups were formed to focus on sectioning the network and identifying potential sampling locations. Instead of using the current RT-PCR method (Heijnen et al., 2024), the sampling procedure was undertaken using another PathoCERT tool named PathoTESTICK, a mobile sensor that offers rapid on-site screening of *Escherichia coli*, in less than 5 minutes (Canciu et al., 2022). For the purposes of this case study, it was assumed that PathoTESTICK can detect wastewater contamination with sufficient sensitivity. Each sampling iteration was estimated to last approximately 20 minutes, which included the time needed for the field team to reach the location, set up the sampling equipment, await the results, and communicate them back.

In response to the potential health risks posed by the contamination, the water utility's emergency response aimed to rapidly locate the source of contamination by deciding on sampling locations, and contain its spread by closing valves, thereby mitigating the infection risk. For the traditional approach, these decisions were based on the response team's expertise, while for the model-based approach, these decisions were guided by the PathoINVEST tool.

5.2.4 Source identification

5.2.4.1 Traditional approach

After discussing with the response team and based on the insights of Figure S5.1 (average flow direction, water age), the incident commander initiated two sampling rounds and designated 11 strategic sampling locations for the field team across the network (Figure 5.1). The incident commander had to wait for the results of the first designated sample before deciding on the next sampling location and the first sample was taken at 10:30 a.m., approximately 1 hour after the complaints. The objective was to leverage the information obtained from the samples (indicating either the presence or absence of contamination) in each round, thereby leading to the exclusion of segments of the network and finally, the identification of the contamination source.

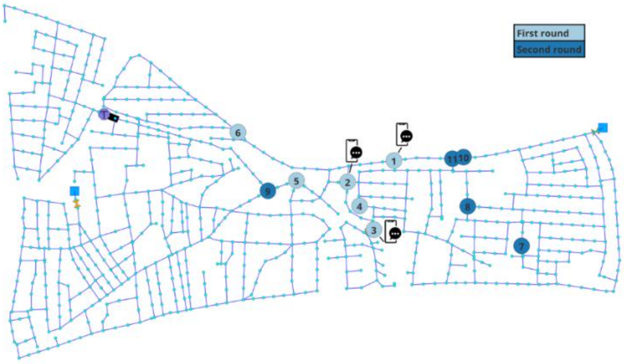


Figure 5. 1 The proposed sampling locations in two rounds (light blue and dark blue), along with the customer complaints (smartphone).

5.2.4.2 Model-based approach

PathoINVEST employs a methodology based on a simplified version of the expanded sampling concept (Eliades and Polycarpou, 2012), for the identification of contamination sources in drinking water networks. Their methodology was based on decision trees, expressing conditional statements such as *if-then-else* rules, to return a sequence of nodes for manual sampling. Our approach involves analyzing potential contamination sources and the strategic selection of sampling locations, each marked with a binary signature dependent on the outcome, positive (1) or negative (0). A positive binary signature at a sampling location signifies that a given contamination source's trajectory has intersected that point, suggesting that any sample retrieved from a field team at this site would yield a positive result. Therefore, the first step for the model-based approach was to identify potential contamination sources. This was achieved by finding any upstream node from the location of the complaints (Figure 5.2). The desired outcome was obtained using an additional feature of the tool, a function $f(\mathcal{G}, s, d)$ able to identify potential upstream contamination nodes in a network. This function finds all nodes within a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ represent the nodes and edges, respectively. Specifically, it locates nodes within an infinite distance d from a specified node s (the complaint node), effectively capturing any node in the network that could contribute to upstream contamination of s.

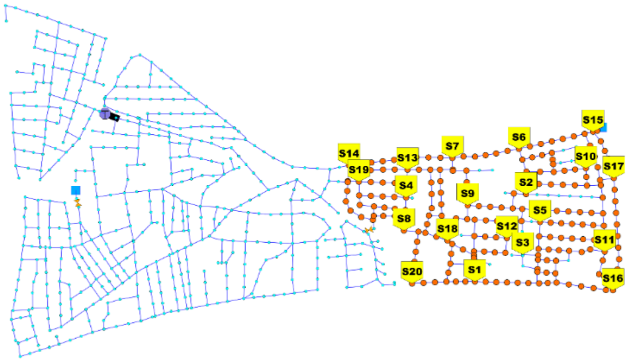


Figure 5. 2 The upstream potential contamination sources (orange circles) using the complaints as a starting point, and the 20 most probable contamination sources (yellow labels).

After identifying and highlighting all the potential sources of upstream contamination, 20 strategic locations were selected as the most probable contamination sources by the response team (Figure 5.2). The next step was to select strategic locations for sampling. All the potential upstream contamination sources from Figure 5.2 were also considered sampling locations (243 nodes in total). Again, the time-to-result of the PathoTESTICK tool was used.

To find the suspected contamination source, the hydraulics and quality dynamics of the DWDN were simulated for 20 scenarios, representing all 20 contamination sources. This enabled the generation of binary signatures for all sampling locations corresponding to the simulated contamination trajectories (Table 5.1). The simulation of the 20 contamination scenarios was approximately 5 minutes. Consequently, after accounting for a 25-minute discussion with the team regarding the 20 strategic locations, the binary signatures and hydraulics were successfully computed by 10:00 a.m. The temporal dynamics of contamination play a critical role in this process. For instance, a node, say N_{220}, may not be associated with a contamination source at 10:00 AM but may become contaminated by 11:00 AM due to the progression of contamination.

Table 5. 1 The simulated binary signatures of a snippet of the sampling locations for each of the 20 contamination sources at 10:00 a.m. Node n_{241} is the selected sampling location based on the entropy results. Red rows indicate the scenarios that are not consistent with the sampling result. The result was positive (1), therefore the red highlighted scenarios can be discarded.

Scenario	n_{220}	n_{221}	n_{222}	n_{223}	n_{224}	n_{225}	n_{228}	n_{229}	n_{230}	n_{232}	n_{235}	n_{238}	n_{239}	n_{240}	n_{241}	n_{244}	n_{245}	n_{248}	n_{249}	n_{250}
S1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
S2	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S4	0	0	0	0	0	1	0	0	0	0	1	1	1	0	1	1	0	0	0	0
S5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
S6	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	0	0	0
S7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S8	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0
S9	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0
S10	0	1	1	1	1	1	0	0	0	0	1	1	1	0	1	1	0	0	0	0
S11	0	1	1	1	1	0	1	1	0	1	1	1	0	0	0	0	0	0	0	0
S12	0	1	1	1	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0
S13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
S15	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
S16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S17	1	1	1	1	0	0	1	0	1	1	0	0	0	1	0	0	1	1	1	1
S18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S19	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	0	0	0
S20	0	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	0	0	0	0

After the generation of binary signatures in all sampling locations for each of the 20 contamination scenarios, the tool indicated the first sampling location. To prioritize the optimal sampling location the tool utilizes the theory of entropy, which measures the uncertainty or unpredictability in a system. In our context, each sampling location can either be contaminated (1) or not contaminated (0). We calculate the probabilities $p1$ and $p0$ based on the frequency of positive and negative outcomes from the 20 contamination simulations. The entropy H at each node v, is computed using the formula:

$$H(v) = -p(0)\log_2\big(p(0)\big) - p(1)\log_2\big(p(1)\big)$$

where $p(0)\log_2\big(p(0)\big)$ and $p(1)\log_2\big(p(1)\big)$ represent the information content or uncertainty when the location is not contaminated and when the location is contaminated respectively.

High entropy indicates a high level of uncertainty about the contamination status at a location, meaning it is equally likely to be contaminated or not. Sampling at locations with high entropy maximizes the informational yield, reducing uncertainty most effectively. By focusing on high-entropy sampling locations, this methodology aims to maximize the informational yield from each sample and thus, efficiently narrow down the possible contamination sources.

After entropy analysis, node n_{241} exhibited the highest entropy, signaling it as the prime candidate for sampling (Table 5.2). In cases where multiple locations displayed equivalent maximum entropy, the selection of the next sampling site was deferred to the discretion of the user.

Table 5. 2 Results of entropy for the selected snippet of sampling locations.

	$\boldsymbol{n_{241}}$	$\boldsymbol{n_{235}}$	$\boldsymbol{n_{223}}$	$\boldsymbol{n_{229}}$	$\boldsymbol{n_{244}}$	$\boldsymbol{n_{238}}$	$\boldsymbol{n_{239}}$	$\boldsymbol{n_{222}}$	$\boldsymbol{n_{224}}$	$\boldsymbol{n_{221}}$	$\boldsymbol{n_{225}}$	$\boldsymbol{n_{228}}$	$\boldsymbol{n_{232}}$	$\boldsymbol{n_{\{240}}$	$\boldsymbol{n_{220}}$	$\boldsymbol{n_{230}}$	$\boldsymbol{n_{245}}$	$\boldsymbol{n_{248}}$	$\boldsymbol{n_{249}}$	$\boldsymbol{n_{250}}$
$\boldsymbol{P(1)}$	0.5	0.55	0.55	0.55	0.55	0.55	0.55	0.6	0.4	0.6	0.35	0.25	0.25	0.25	0.2	0.18	0.15	0.15	0.15	0.15
$\boldsymbol{P(0)}$	0.5	0.45	0.45	0.45	0.45	0.45	0.45	0.4	0.6	0.4	0.65	0.75	0.75	0.75	0.8	0.85	0.85	0.85	0.85	0.85
$\boldsymbol{H}$	1	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.97	0.97	0.93	0.81	0.81	0.81	0.72	0.6	0.6	0.6	0.6	0.6

The field team was instructed to take the first sample in the designated location (node n_{241}) at 10:00 a.m. The result was positive and that allowed for the exclusion of simulated contamination scenarios that do not align with this outcome (Table 5.1). This iterative process continued until the set of potential contamination sources was sufficiently narrowed, facilitating the identification of the actual source of contamination. The remaining procedure (identification of the sampling node with the highest entropy and exclusion of simulated contamination scenarios can be found in the supplementary material).

5.2.5 Valve manipulation

Based on the network topology and discussions with the team, the incident commander proceeded to the closure of valves at strategic locations. The first mitigation action (for both traditional and model-based approaches) was the closure of valves V_1 and V_2, located near the complaint site (Figure 5.4b), at 10:40 a.m. and 10:45 a.m., respectively (*Action 1*). This action was aimed at preventing further contamination spread to the network's western part even though the contamination source was not found yet. This was considered an appropriate action because it would keep the water supply intact for all customers while separating the network into two parts. Each part would be supplied by its own water source, preventing contamination from moving between these two sections of the network. As soon as the contamination source was identified, the second mitigation action from the water utility (again for both traditional and model-based approaches) was to immediately close the pipes surrounding the source to prevent further spreading (*Action 2*).

5.2.5.1 Traditional approach

Although the traditional *Action 2* stopped the contamination source at 2:00 p.m., pathogens were still spreading in the eastern part of the network. Therefore, the water utility decided to implement *Action 3*, which involved closing an additional three valves. Subsequently, valves V_3, V_4, and V_5 were closed at 2:20 p.m., 2:25 p.m., and 2:30 p.m. respectively to contain the spread in the network's eastern part. At this point, it was time to run a model simulation to evaluate the performance of the traditional approach. The identified contamination source was incorporated to calculate the risk of infection (defined as the percentage of people being infected) and assess whether the health impact was mitigated. This simulation specifically factored in which valves to close (spatial resolution) and the timing of their closure (temporal resolution).

5.2.5.2 Model-based approach

For the model-based approach, *Actions 1* and *2* are identical to the traditional approach. The model-based *Action 2* occurred earlier at 11:30 a.m. since the source was identified earlier. This time the additional closure of three valves (model-based *Action 3*), was suggested by the PathoINVEST tool. The PathoINVEST methodology on valve manipulation expands the work of Moghaddam et al. (2022) focusing particularly on the application of Particle Swarm Optimization (PSO) for the strategic closure of pipes within a DWDN to mitigate a contamination event. Diverging from their original model, which includes both pipe closures and hydrant activations, we refined our approach to solely concentrate on pipe closures with a primary objective of minimizing the infection risk, rather than minimizing the number of contaminated nodes. The PSO algorithm is fine-tuned to identify optimal pipe closure strategies that effectively reduce the infection risk once the contamination source is known. By simulating the movements of particles within a swarm, each particle represents a potential solution. Through iterative refinements and adjustments to particle positions and velocities, our modified PSO model searches for the most effective configuration, emphasizing the minimization of infection risk while also incorporating a penalty function to ensure system pressures are maintained above critical thresholds. Another modification compared to the original work of Moghaddam et al. (2022) is that we restricted the PSO algorithm to recommend closing only three pipes in response to real-time contamination events. This limitation aims to increase realism and feasibility for water utilities, recognizing the practical challenges of implementing extensive infrastructure modifications during an emergency. Moreover, by narrowing the potential actions to three pipe closures, we significantly reduce the solution space, which, in turn, decreases computational demands and time. This adjustment not only aligns with the operational capabilities of water utilities but also reduces the significant computational resources typically required by evolutionary algorithms, including the need for advanced processing power and significant memory capacity to efficiently manage iterative optimizations. After the PathoINVEST feature recommended the three most suitable valves for closure, another simulation was conducted with the new valve settings to assess the risk of infection.

5.2.6 Valve settings uncertainty

In hydraulic modeling, the fidelity of the network representation is very important. When the aim is to model contaminations, predict realistic outcomes, find the contamination source, and suggest adequate mitigation actions, there must be an accurate representation of the most current network configurations, including valve settings. Inaccurate or outdated models can produce misleading results, making them unreliable for decision-making processes. This unreliability is particularly pronounced in scenarios where valve settings vary significantly between the actual network and the model (e.g., a valve is open in reality but closed in the model). These discrepancies can drastically change the simulated hydraulics and flow directions, thus providing inaccurate results. Previous studies highlight that there can be discrepancies of up to 0.7% between the model and actual valve settings (Mesman et al., 2016).

To explore the potential effect of this inherent uncertainty in valve settings, we developed an approach involving the generation of multiple hydraulic profiles. Specifically, we created 1000 distinct hydraulic profiles for L-Town, each incorporating a different 0.7% variation in closed valve settings (6 out of 905 valves). These profiles were then used to simulate a contamination scenario with a known contamination source. This approach generated 1000 unique uncertainty scenarios ensuring that both hydraulic and pressure requirements of the system were met. For each scenario, we calculated the risk of infection, allowing us to quantitatively assess the impact of valve uncertainty on the overall risk of infection and compare them with the base scenario (the scenario that was used in the pilot case study without any mitigation action). Furthermore, we evaluated the impact of valve uncertainty, as represented by the 1000 uncertainty scenarios, on the efficacy of the model-based approach in identifying the contamination source using the expanded sampling, as detailed in Chapter 5.2.4.2.

5.3 Results and Discussion

Figure 5.3 shows the contamination source (S_7) and the potential contamination propagation to its full extent by midnight (14.5 hours after the complaints were received) without any action by the water utility. The figure also displays the sampling times and results for both approaches (traditional and model-based).

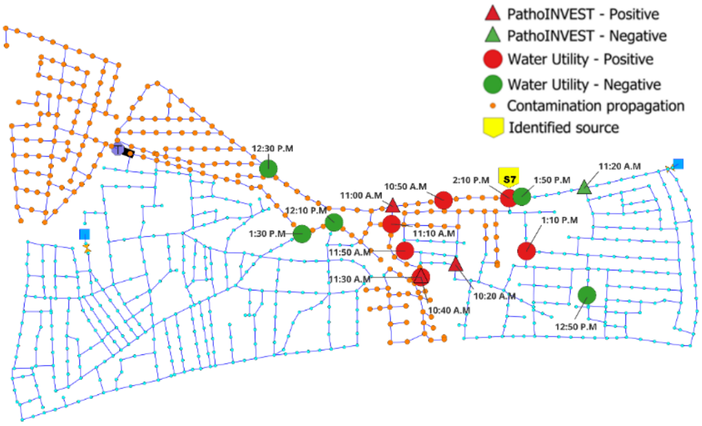


Figure 5. 3 The identified source, contamination propagation, and sampling results for both traditional and model-based approaches.

5.3.1 Source identification

Table 5.3 compares the traditional and the model-based approach, detailing the number of samples and the time needed to find the contamination source. In addition, it assesses the estimated infection risk starting from the onset of contamination (8:00 a.m.) until the moment the contamination source was finally identified both with and without the closing of two valves (*Action 1*). As a point of reference, we also calculated the infection risk using the RT-PCR method (4 hours) instead of the PathoTESTICK, which yields results in 20 minutes. Although considerably faster than the culture-based approach, the RT-PCR approach would still mean that the field team would face a 4-hour wait for each round of results. Since two sampling rounds were performed, this would result in a total waiting time of at least 8 hours.

The model-based approach is more efficient and rapid in locating the source, as it takes only 4 samples and 1.3 hours, compared to 11 samples and 3.7 and 8 hours for the traditional approach, using the PathoTESTICK and RT-PCR respectively. Calculating the risk of infection until the time the contamination source was identified (and any potential mitigation measures could have started) for all three approaches we see that the model-based approach is again better here as the risk of infection is 12% compared to 20% and 27% for the traditional approach with the PathoTESTICK and RT-PCR respectively. Finally, a key observation is the (moderate) reduction of the health impact of the contamination by the water utility's rapid action on closing the first two valves (*Action 1*) compared to the base scenario (Table 5.3 and Figure 5.4a).

Table 5. 3 Comparison between traditional and model-based approaches to find the source of contamination.

	No of samples	**Time to find the source (hours)**	**Infection risk with Action 1**	**Infection risk w/o Action 1**
Traditional (RT-PRC)	11	8	27 %	33 %
Traditional (PathoTESTICK)	11	3.7	20 %	22 %
Model-based (PathoTESTICK)	4	1.3	12 %	12 %

5.3.2 Valve manipulation

5.3.2.1 *Traditional approach*

For the traditional approach, the closure of two valves identified as V_1 and V_2 (*Action 1*), at 10:40 a.m. and 10:45 a.m. respectively, effectively reduced the infection risk, as shown in Figure 5.4a. However, since the contamination was initiated at 8:00 a.m., a certain extent of contaminant spread had occurred before *Action 1* was implemented. As a result, the valve closures, while appropriate, were late, leading to considerable contamination in L-Town's western region. Additionally, the traditional *Action 3* inadvertently redirected the contamination plume, reaching previously unaffected zones. This misdirection increased the potential risk of infection as can be seen in Figure 5.4a.

5.3.2.2 *Model-based approach*

Utilizing the modified PSO algorithm, the model-based approach identified the optimal three valves, V_6, V_7, and V_8, for closure after identifying the contamination source at 11:20 a.m., as shown in Figure 5.4b. These pipes were closed around noon, considering the travel time for the field team and the time needed for valve closure. As shown in Figure 5.4a, the model-based *Action 3* reduced the infection risk when compared to the base scenario as well as the traditional *Action 3*. However, the infection risk remains relatively high. Due to the timing of this action, those at risk had likely already been exposed.

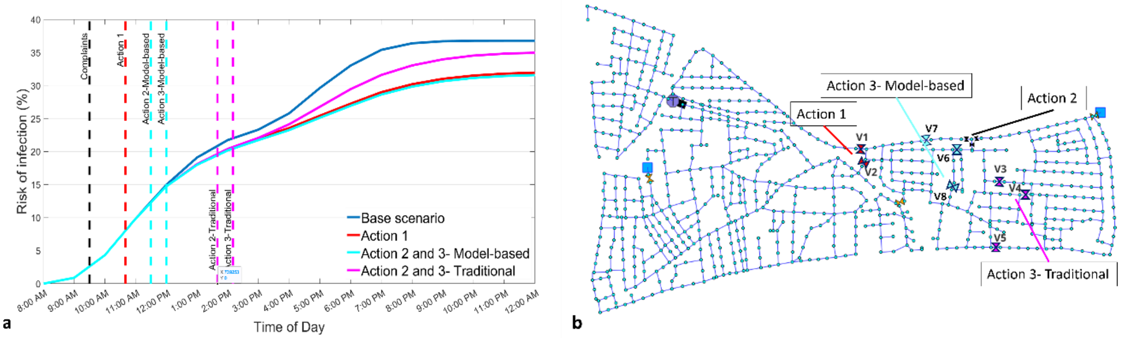


Figure 5. 4 a) Risk of infection comparison between the base scenario, and the mitigation Actions 1-3. b) Closure of valves (Actions) by the water utility and PathoINVEST.

5.3.3 Importance of rapid response

Figure 4a shows that the infection risk although reduced, remained high regardless of mitigation actions from the water utility. This indicates that to significantly minimize the infection risk, interventions should have been earlier. A re-evaluation of the event timeline, featuring an alternative simulation of the same contamination scenario where actions were initiated one hour earlier, sheds light on the importance of rapid response. In this revised scenario, valves V_1 and V_2 were closed at 9:40 a.m., immediately following the customer complaints, as a `no-regret' preventative mitigation action. Using PathoINVEST to select the sampling locations also moved subsequent mitigation *Actions 2* and *3* to one hour earlier. This quicker response led to a significant decrease in the infection risk (a 17%-point reduction in total infection risk), indicating that immediate valve closure after contamination detection (or even suspicion) can significantly contain the contamination (Figure 5.5). This highlights the importance of strategic sensor placement, rapid response actions (automated valve closures), and sensors that offer screening of fecal contamination during such emergencies.

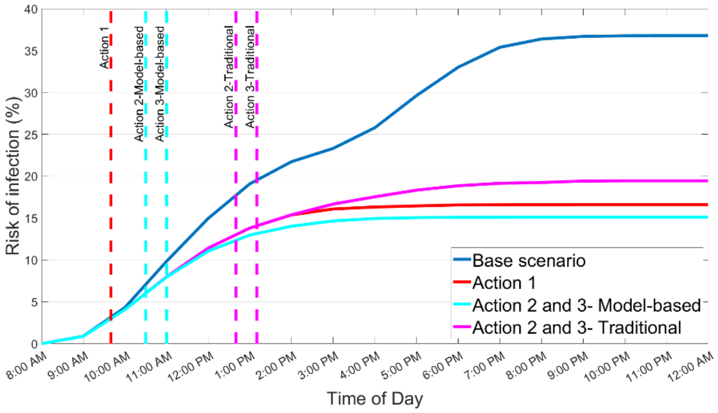


Figure 5. 5 Risk of infection comparison between the base scenario and the mitigation Actions 1-3 in the alternative contamination simulation where all actions are advanced by an hour.

5.3.4 Valve settings uncertainty

5.3.4.1 Infection risk uncertainty

To evaluate the potential impact of valve settings uncertainty, Figure 5.6a demonstrates the infection risk across 1000 uncertainty scenarios with the green line representing the base scenario. The 0.7% uncertainty in valve settings yields variability in the projected infection risk. Figure 5.6b shows the infection risk difference of the 1000 uncertainty scenarios from the base scenario. The histogram indicates that about 80% of the uncertainty scenarios are within the range of -5% to +5% from the base scenario infection risk. Also, the majority of the 1000 uncertainty scenarios cluster around the base scenario. This suggests that modeling with a 0.7% uncertainty in valve settings generally does not result in significantly different or underestimated infection risk outcomes.

The scenarios with significantly lower infection risk (below 10%) are observed when one or more pipes downstream of the eastern reservoir are actually closed, or in cases where pipes located immediately downstream of the contamination source are closed. These scenarios effectively limit the spread of contamination, thereby significantly reducing the infection risk.

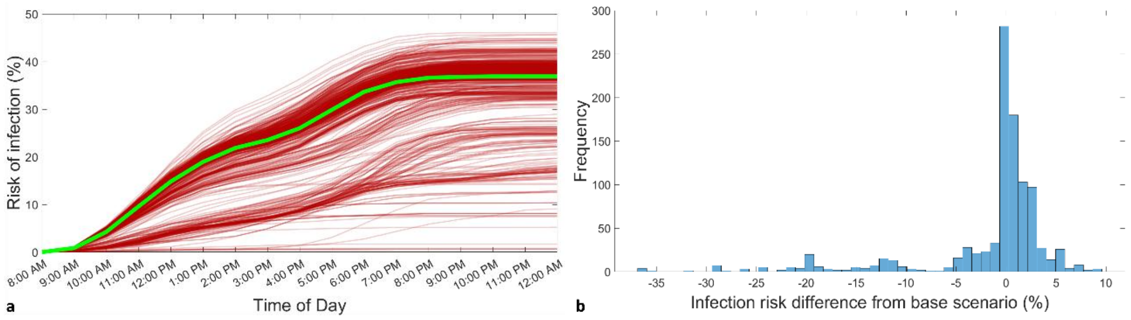


Figure 5. 6 a) Risk of infection for the 1000 uncertainty scenarios and the base scenario (green line). b) Histogram showing the difference between the 1000 uncertainty scenarios in infection risk compared to the base scenario.

5.3.4.2 Source identification uncertainty

Valve settings uncertainty may also impact the reliability of the PathoINVEST contamination source identification feature. Figure 5.7 showcases how a 0.7% uncertainty in valve settings influenced the performance of the model-based approach in identifying the contamination source. This analysis was conducted for each of the 1000 uncertainty scenarios, following the expanded sampling methodology detailed in Chapter 5.2.4.2 for the 20 potential contamination sources. The results indicate that, despite the valve settings uncertainty, the model successfully identified the exact source of contamination, S_7, in 57% of the uncertainty scenarios. The next most frequently identified source was S_6 with 18%, which is adjacent to the exact source, S_7. This implies that even if PathoINVEST incorrectly pinpoints S_6 as the contamination source, it would still guide the investigation toward the correct vicinity for further inspections. Considering this, the effectiveness of PathoINVEST source identification feature in correctly identifying the source of contamination stands at 75%.

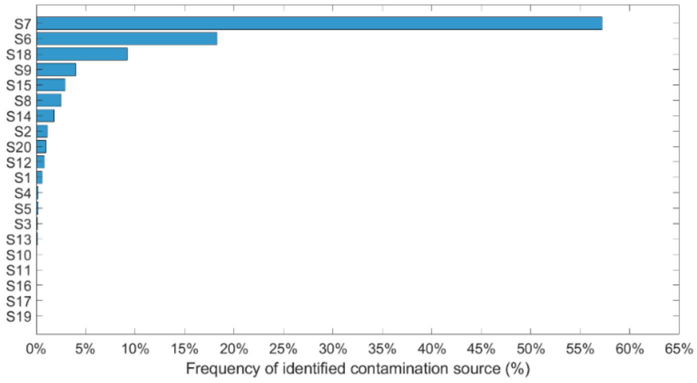


Figure 5. 7 Frequency of the 1000 uncertainty scenarios in identifying the contamination source using PathoINVEST.

5.3.4.3 Valve manipulation under valve settings uncertainty

The influence of valve settings uncertainty on valve manipulation was evaluated only for *Action 1*. Figure 5.8 shows the mean (derived from the 1000 uncertainty scenarios) total infection risk associated with 10 of the potential 20 contamination sources, comparing the base scenario (*no action*) with the original closure of two pipes (V_1 and V_2) by the water utility (*Action 1*). This comparison demonstrates that, regardless of the contamination source and its accurate identification by the PathoINVEST source identification feature (due to any potential valve settings uncertainty), the strategic closure of two pipes (*Action 1*) consistently reduced the relative risk of infection, ranging from 23.8% (S_7) to 62.4% (S_8) reduction. This outcome highlights the effectiveness of strategic valve closure during a contamination event.

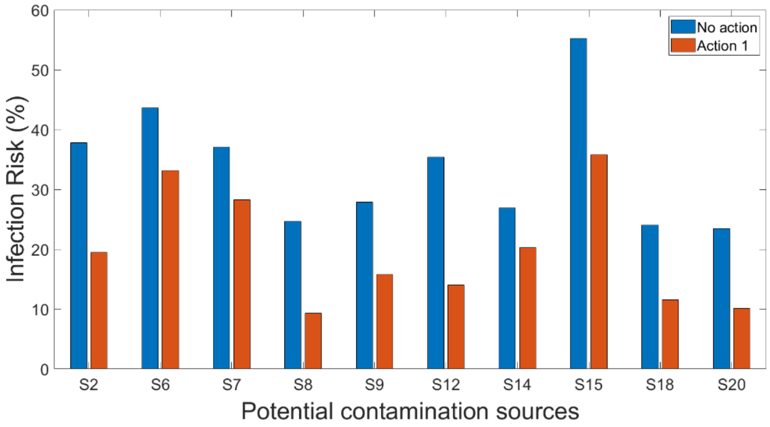


Figure 5. 8 Comparison of the mean infection risk (derived from 1000 uncertainty scenarios) between the base scenario (no action) and Action 1 for the most frequently identified contamination sources by PathoINVEST.

5.3.5 Approach assessment

Assessing traditional and model-based approaches to managing such emergencies allows identifying their respective strengths and weaknesses.

5.3.5.1 *Traditional approach*

The decisions made in the traditional approach rely heavily on the experience and intuition of the response team whereas the need for a good understanding of the network characteristics is imperative. This approach, while it doesn't possess the sophistication of advanced technologies, offers reliability and independence from modeling tools, which is important in situations where such technologies may be unavailable or slow. We saw that the response team's quick decision to close the two valves at the beginning of the event (*Action 1*), led to a reduction in infection risk. However, this approach tends to be more time intensive. This is demonstrated by the 11 samples required to identify the contamination source, compared to just 4 from the model-based approach, resulting in an additional 2 hours and 20 minutes. During this time, there was an additional 8%-point infection risk. It can also sometimes lead to mistakes, such as the wrong closure of the remaining three valves (traditional *Action 3*) which resulted in a 3%-point higher infection risk compared to both *Action 1* and model-based *Action 3*. The lack of predictive capabilities inherent in this approach often results in slower decision-making processes, extensive response times, and mitigation measures of diminished efficacy. This issue becomes even more pronounced when considering current detection and sampling methods (such as RT-PCR), which require considerably more time to deliver results, further slowing down the decision-making process during emergencies.

5.3.5.2 *Model-based approach*

Having an on-site mobile device capable of detecting pathogens in water with 20 minutes time-to-result is extremely valuable and important during emergencies. The model-based approach demonstrates significant improvements in response speed since it takes 2 hours and 20 minutes less to find the source in the presented case study. Being able to simulate multiple real-time contamination events, propose sampling locations for source identification, and suggest valve closures to mitigate the event offers valuable insights to a water utility resulting in more effective decision-making. Nonetheless, the model-based approach also has limitations. First of all, in this work we focus on an unchlorinated network, which means the E. coli sensor is effective. However, in a chlorinated network, E. coli would be immediately inactivated, making the sensor essentially useless. This is a limitation that highlights the need for pathogen-specific sensors for broader applicability. Moreover, the accuracy of a model is heavily dependent on the quality and current state of network data. An inaccurate or outdated network can compromise the model's outputs, leading to potentially flawed decision-making by the response team. Additionally, these tools require constant updates and maintenance, which requires commitment of resources. The deterministic nature of models also presents a limitation, as they might not fully capture the complexity and variability of real-world scenarios. While the application of PSO for valve manipulation demonstrates an efficient outcome, it may not represent the optimal approach. Alternative metaheuristic algorithms (e.g., multiobjective evolutionary algorithms) could potentially offer improved results for valve closure (Nouiri 2017; Quintiliani et al. 2019). The constraint within our PSO application, specifically to close only three valves, introduces another layer of limitation. However, the primary aim of this work was to investigate the added value of using modeling tools and rapid sensors during emergencies. Further improvements in the sensor and the presented methodology (source identification, valve manipulation) can enhance the benefits of using mobile devices for on-site contamination screening and models in the decision-making of crisis scenarios.

5.4 Conclusions

This study compared traditional and model-based approaches in managing DWDN contamination events, revealing several key insights:

1. The water utility's rapid response action to close the first two valves (traditional approach), despite not knowing the contamination source, effectively prevented the spread of contamination and reduced the health impact.
2. The model-based approach was shown to be more efficient than the traditional approach in identifying the source of contamination (1.3 versus 3.7 hours), requiring fewer samples (4 versus 11) and resulting in lower infection risk by the time the source was identified (12% versus 20%) in this case study.
3. The model-based approach was more effective in finding the best valves to close in the network since it resulted in a 3%-point infection risk reduction.
4. Having up-to-date valve settings in the DWDN schematization is important to provide reliable results on source identification. Discrepancies between the actual network and the model can lead to inaccurate infection risk estimates when using modeling tools to support decision-making.
5. Rapid actions and decision-making are crucial upon detecting contamination in the DWDN, as a 1-hour faster response from the water utility can lead to a 17%-point reduction in total infection risk. One example of such rapid actions is the use of mobile rapid testing devices for on-site contamination screening, as they deliver immediate results and enable quicker responses.

This study underscored the potential advantages of integrating modeling and sensor tools for managing DWDN contamination events. It demonstrated improvements in efficiency and speed of model-based approaches, dependent on the network model's accuracy and effective management of uncertainties, such as valve settings. Furthermore, the study highlighted the essential value of traditional knowledge and human intuition in emergency responses, illustrating how quick expert decisions remain critical. The deployment of mobile devices for rapid on-site contamination screening represents a significant advancement, facilitating immediate response actions. By combining model-based strategies with traditional expert insights, our approach provides a robust framework for improving water contamination management and decision-making processes, thus ensuring public health during emergencies. The development of real-time modeling tools such as PathoINVEST further exemplifies this approach, showing great potential for promoting operational improvements in drinking water crisis management. Our case study showed that when a contamination event unfolds, the window of opportunity for meaningful interventions is small (within a few hours), while the risk of infection can be quite high since many people can be exposed to pathogens.

Data availability statement

Some or all data, models, or code generated or used during the study are available in a repository or online in accordance with funder data retention policies.

https://github.com/KIOS-Research/PathoINVEST-WDSA-CCWI-2022

Acknowledgements

This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 883484 (PathoCERT), the European Research Council (ERC), under the ERC Synergy grant agreement No. 951424 (Water Futures), and was supported by the European Union's Horizon 2020 Teaming programme under grant agreement No. 739551 (KIOS CoE), and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

Supplementary material

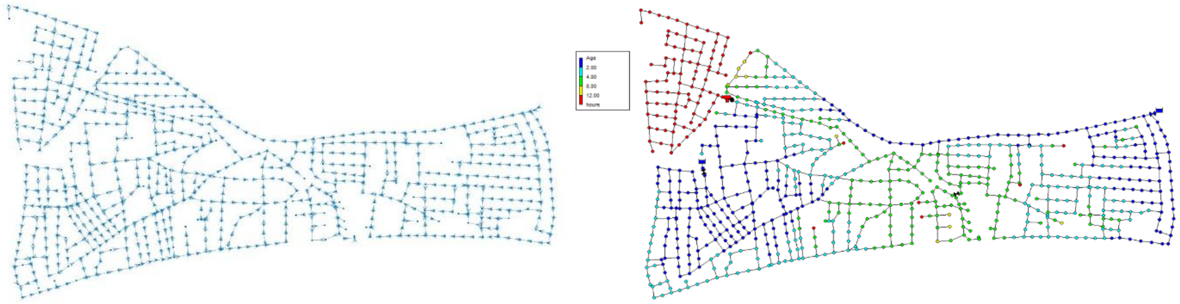


Figure S5. 1 Daily average flows (left) and water age (right) of L-Town.

Table S5.1 The selected sampling location (green color) based on the entropy results for the second iteration. Red rows indicate the scenarios that are not consistent with the sampling result. The result was positive (1) therefore the red highlighted scenarios can be discarded.

Scenario	n_{220}	n_{221}	n_{222}	n_{223}	n_{224}	n_{225}	n_{228}	n_{229}	n_{230}	n_{232}	n_{235}	n_{238}	n_{239}	n_{240}	n_{241}	n_{244}	n_{245}	n_{248}	n_{249}	n_{250}
S2	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S4	0	0	0	0	0	1	0	0	0	0	1	1	1	0	1	1	0	0	0	0
S6	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	0	0	0
S7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S8	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0
S9	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0
S10	0	1	1	1	1	1	0	0	0	0	1	1	1	0	1	1	0	0	0	0
S14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
S19	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	0	0	0
S20	0	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	0	0	0	0

Table S5.2 The selected sampling location (green color) based on the entropy results for the third iteration. Red rows indicate the scenarios that are not consistent with the sampling result. The result was positive (1) therefore the red highlighted scenarios can be discarded.

Scenario	n_{220}	n_{221}	n_{222}	n_{223}	n_{224}	n_{225}	n_{228}	n_{229}	n_{230}	n_{232}	n_{235}	n_{238}	n_{239}	n_{240}	n_{241}	n_{244}	n_{245}	n_{248}	n_{249}	n_{250}
S2	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S6	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	0	0	0
S7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S8	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0
S19	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	0	0	0

Table S5.3 The selected sampling location (green color) based on the entropy results for the fourth iteration this time at 11:00 a.m. The red row indicates the scenarios that are not consistent with the sampling result. The result was negative (0) therefore the red highlighted scenario can be discarded.

Scenario	n_{251}	n_{252}	n_{255}	n_{256}	n_{257}	n_{258}	n_{259}	n_{260}	n_{261}	n_{262}	n_{263}	n_{264}	n_{265}	n_{266}	n_{267}	n_{268}	n_{269}	n_{270}	n_{271}	n_{272}
S2	1	1	1	0	1	1	0	0	0	1	1	1	0	1	1	0	1	0	1	0
S7	1	1	1	0	1	1	0	0	0	1	1	1	0	1	1	0	1	0	0	0

References

Afshar, A. & Najafi, E. 2014. Consequence management of chemical intrusion in water distribution networks under inexact scenarios. *Journal of Hydroinformatics*, *16*(1), 178-188.

Ailes, E. Budge, P. Shankar, M. Collier, S. Brinton, W. Cronquist, A. ... & Brunkard, J. M. 2013. Economic and health impacts associated with a Salmonella typhimurium drinking water outbreak− Alamosa, CO, 2008. *PloS one*, *8*(3), e57439.

Anadu, E. C. & Harding, A. K. 2000. Risk perception and bottled water use. *Journal-American Water Works Association*, *92*(11), 82-92.

Arnone, R. D. & Perdek Walling, J. 2007. Waterborne pathogens in urban watersheds. *Journal of Water and Health*, *5*(1), 149-162.

Blackburn, B. G. Craun, G. F. Yoder, J. S. Hill, V. Calderon, R. L. Chen, N. ... & Beach, M. J. 2004. Surveillance for waterborne-disease outbreaks associated with drinking water—United States, 2001–2002. *MMWR Surveill Summ*, *53*(8), 23-45.

Blokker, M. Smeets, P. & Medema, G. 2018. Quantitative microbial risk assessment of repairs of the drinking water distribution system. *Microbial Risk Analysis*, *8*, 22-31.

Canciu, A. Cernat, A. Tertis, M. Botarca, S. Bordea, M. A. Wang, J. & Cristea, C. 2022. Proof of Concept for the Detection with Custom Printed Electrodes of Enterobactin as a Marker of Escherichia coli. *International Journal of Molecular Sciences*, *23*(17), 9884.

Cann, K. F. Thomas, D. R. Salmon, R. L. Wyn-Jones, A. P. & Kay, D. 2013. Extreme water-related weather events and waterborne disease. *Epidemiology & Infection*, *141*(4), 671-686.

Chyzheuskaya, A. Cormican, M. Srivinas, R. O'Donovan, D. Prendergast, M. O'Donoghue, C. & Morris, D. 2017. Economic assessment of waterborne outbreak of cryptosporidiosis. *Emerging Infectious Diseases*, *23*(10), 1650.

Cominola, A. Giuliani, M. Castelletti, A. Rosenberg, D. E. & Abdallah, A. M. 2018. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. *Environmental Modeling & Software*, *102*, 199-212.

Corso, P. S. Kramer, M. H. Blair, K. A. Addiss, D. G. Davis, J. P. & Haddix, A. C. 2003. Costs of illness in the 1993 waterborne Cryptosporidium outbreak, Milwaukee, Wisconsin. *Emerging infectious diseases*, *9*(4), 426.

Eliades, D. G. & Polycarpou, M. M. 2012. Water contamination impact evaluation and source-area isolation using decision trees. *Journal of Water Resources Planning and Management*, *138*(5), 562-570.

Eliades, D. G. Kyriakou, M. Vrachimis, S. & Polycarpou, M. M. 2016, November. EPANET-MATLAB toolkit: An open-source software for interfacing EPANET with MATLAB. In *Proc. 14th international conference on computing and control for the water industry (ccwi)* (Vol. 8). The Netherlands.

Eliades, D. G. Vrachimis, S. G. Moghaddam, A. Tzortzis, I. & Polycarpou, M. M. 2023. Contamination event diagnosis in drinking water networks: A review. *Annual Reviews in Control*.

Erickson, T. B. Brooks, J. Nilles, E. J. Pham, P. N. & Vinck, P. 2019. Environmental health effects attributed to toxic and infectious agents following hurricanes, cyclones, flash floods and major hydrometeorological events. *Journal of Toxicology and Environmental Health, Part B*, *22*(5-6), 157-171.

Fewtrell, L. Kay, D. Watkins, J. Davies, C. & Francis, C. 2011. The microbiology of urban UK floodwaters and a quantitative microbial risk assessment of flooding and gastrointestinal illness. *Journal of flood risk management*, *4*(2), 77-87.

Gude, V. G. & Muire, P. J. 2021. Preparing for outbreaks–implications for resilient water utility operations and services. *Sustainable Cities and Society*, *64*, 102558.

Heijnen, L. de Vries, H. J. van Pelt, G. Stroobach, E. Atsma, A. Vranken, J. ... & Medema, G. 2024. Qualitative detection of E. coli in distributed drinking water using real-time reverse transcription PCR targeting 16S rRNA: validation and practical experiences. *Water Research*, 121843.

Hrudey, S. E. & Hrudey, E. J. 2004. *Safe drinking water*. IWA publishing.

Kunz, J. M. 2024. Surveillance of Waterborne Disease Outbreaks Associated with Drinking Water—United States, 2015–2020. *MMWR. Surveillance Summaries*, *73*.

Laine, J. Huovinen, E. Virtanen, M. J. Snellman, M. Lumio, J. Ruutu, P. ... & Pirkanmaa Waterborne Outbreak Study Group. 2011. An extensive gastroenteritis outbreak after drinking-water contamination by sewage effluent, Finland. *Epidemiology & Infection*, *139*(7), 1105-1113.

Laursen, E. Mygind, O. Rasmussen, B. & Rønne, T. 1994. Gastroenteritis: a waterborne outbreak affecting 1600 people in a small Danish town. *Journal of Epidemiology & Community Health*, *48*(5), 453-458.

Lendowski, L. Färber, H. Holy, A. Darius, A. Ehrich, B. Wippermann, C. ... & Exner, M. 2015. Accidental contamination of a German town's drinking water with sodium hydroxide. *International journal of hygiene and environmental health*, *218*(3), 366-369.

Mesman, G. A. M. Noord-Holland, P. W. & Evides, N. V. 2016. Automatische registratie afsluiterstanden. *KWR2016*, *50*.

Moghaddam, A. Afsharnia, M. Mokhtari, M. & Peirovi-Minaee, R. 2022. Management and health risk assessment of chemical contamination events in water distribution systems using PSO. *Environmental Monitoring and Assessment*, *194*(5), 362.

Nouiri, I. 2017. Optimal design and management of chlorination in drinking water networks: a multi-objective approach using genetic algorithms and the Pareto optimality concept. *Applied Water Science*, *7*, 3527-3538.

Paraskevopoulos, S. Vrachimis, S. Kyriakou, M. Pavlou, P. Kouzapas, D. Milis, G. Smeets, P. Eliades, D. Medema, G. Polycarpou, M. & Panayiotou, C. 2022. PathoINVEST: Pathogen contamination investigations during emergencies. In 2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Indsustry. Universitat Politecnica de Valencia. https://doi.org/10.4995/WDSA-CCWI2022.2022.14799

Poulin, A. Mailhot, A. Grondin, P. Delorme, L. Periche, N. & Villeneuve, J. P. 2008. Heuristic approach for operational response to drinking water contamination. *Journal of Water Resources Planning and Management*, *134*(5), 457-465.

Quintiliani, C. Marquez-Calvo, O. Alfonso, L. Di Cristo, C. Leopardi, A. Solomatine, D. P. & De Marinis, G. 2019. Multiobjective valve management optimization formulations for water quality enhancement in water distribution networks. *Journal of Water Resources Planning and Management*, *145*(12), 04019061.

United States Environmental Protection Agency. 2003. Response protocol toolbox: Planning for and responding to drinking water contamination threats and incidents—Overview and application.

Vrachimis, S. G. Eliades, D. G. Taormina, R. Kapelan, Z. Ostfeld, A. Liu, S. ... & Polycarpou, M. M. 2022. Battle of the leakage detection and isolation methods. *Journal of Water Resources Planning and Management*, *148*(12), 04022068.

Winston, G. Lerman, S. Goldberger, S. Collins, M. & Leventhal, A. 2003. A tap water turbidity crisis in Tel Aviv, Israel, due to technical failure: toxicological and risk management issues. *International Journal of Hygiene and Environmental Health*, *206*(3), 193-200.

Chapter 6

6. Concluding Chapter

6.1 General remarks

The overall objective of this thesis was to address the growing challenge of waterborne pathogen contamination in DWDNs and develop applications that can enhance decision-making and immediate actions in such emergencies. DWDNs are vital for providing safe drinking water but they are also vulnerable to contamination due to various factors such as infrastructure failures, natural disasters, and even malicious attacks. The thesis developed and evaluated innovative approaches for AI-based information extraction, and advanced modeling tools to help water utilities answer critical questions during an emergency response. Specifically, the questions were related to rapid access to up-to-date information on pathogen characteristics, historical information on contamination events, realistic representation of hydraulics and quality dynamics, predicted health impact, and enhanced decision-making for management and mitigation of a pathogen contamination event. This Chapter summarizes the key findings of this research through answering the research questions, discusses the implications for water utilities, provides examples of practical applications of the proposed methodologies and offers perspectives for future research.

6.2 Answering the research questions

The research questions and objectives outlined in Chapter 1 were systematically addressed through the research studies as presented in Chapters 2-5. Each Chapter contributed unique insights to the overarching challenge of rapid response to contamination events in the DWDN.

6.2.1 First Objective

The research question with title “**Is it feasible to use an automated approach to extract information on waterborne pathogens from the literature?”** was addressed in Chapters 2 and 3, each chapter focusing on a specific sub-question.

Sub-question 1

“Can we use an AI-based model to extract accurate waterborne pathogens-associated information from the scientific literature?”

Chapter 2 developed and evaluated an AI-based methodology to extract specific pathogen information from the scientific literature. By leveraging NLP and DL techniques, the study demonstrated that AI can indeed be used to rapidly extract critical information from scientific publications and be a useful approach for water utilities, enabling faster and more informed decision-making during the early stages of a contamination. The study highlighted the potential of AI to reduce reliance on time-consuming manual literature reviews in environmental microbiology scientific publications.

Sub-question 2

“How effectively do LLMs answer questions related to health risks and mitigation measures about drinking water contamination events, found in the scientific literature?”

Chapter 3 assessed the capacities of open-source LLMs in processing and interpreting complex scientific publications on pathogen contamination events of drinking water through a question-answering task. The evaluation metrics focused on the ability of these models to correctly predict an answer with and without content, as well as their adaptability to answer difficult questions. The results indicated that while LLMs can significantly enhance the information extraction process and provide rapid responses when the question is straightforward, they also face challenges such as hallucinations and generating incorrect answers when questions are complex or difficult even for humans to interpret correctly. Hybrid models that can combine strengths from multiple LLMs could further enhance IE activities. The study demonstrated that when it comes to rapid information extraction approaches during emergencies, LLMs are indeed a proven and effective solution having an average F1 score ranging from 81% to 87%.

6.2.2 Second Objective

The first research question **“What is the health impact of a wastewater contamination in the DWDN considering factors such as chlorine disinfection, hydraulic uncertainty, initial pathogen concentration, pathogen inactivation rate, contamination duration, and contamination location?**” was addressed in Chapter 4. To answer this question, the BeWaRE benchmark testbed was presented, integrating all current relevant knowledge on pathogen transport and fate, bulk and wall chlorine decay, fast and slow chlorine reactions with CRA, CRA degradation, stochastic water demands, hydraulic uncertainty, and individual consumption patterns to calculate pathogen exposure and infection risk following the steps of QMRA. A wastewater contamination in different locations in a chlorinated and non-chlorinated network was simulated using three pathogens: *Campylobacter*, enterovirus, and *Cryptosporidium*. The findings in this Chapter demonstrated that while chlorination aids mitigation, large contaminations can still lead to infections due to chlorine resistance (for Cryptosporidium) and chlorine depletion at the contamination point and the distal part of the DWDN. Moreover, the contamination location and duration, along with varying levels of pathogen susceptibility to chlorine, significantly influence the infection risk, while the response window to reduce the health impact a is short; in these scenarios 5-10 hours post-contamination. The study provided a novel approach to assessing health risks, that can be translated to a valuable decision-support tool for water utilities to optimize their response during emergencies.

The second research question **“What is the added value of using modeling tools to support decision making during wastewater contamination events in the DWDN?”** was addressed in Chapter 5. In this Chapter, a traditional approach was compared with a model-based approach (using PathoINVEST which incorporates the BeWaRE benchmark testbed) for managing a contamination event in the DWDN. Using a case study, the research demonstrated that the model-based approach offers substantial improvements in response times, number of samples needed, and efficiency in terms of contamination source identification. Moreover, regarding mitigation measures, again the model-based approach was better since it provided an optimized valve closure sequence, yielding a lower infection risk than the traditional approach. Finally, it was shown that an up-to-date representation of the DWDN’s valve settings is important to support the decision-making of a water utility.

6.3 Scientific contributions of this Thesis

The scientific contributions of this thesis can be summarized as follows:

- **AI-Based Information Extraction:** An AI-based model was developed and evaluated for extracting both qualitative and quantitative pathogen characteristics from scientific publications. The results showed a satisfactory performance, indicating that water utilities can leverage such models for rapid access to critical information during the early stages of a pathogen contamination event.
- **Assessment of Large Language Models (LLMs):** A systematic evaluation of various open-source LLMs was performed for their ability to perform a question-answering task related to contamination events of DW. The results showed that LLMs have the potential to support decision-making during emergencies since Llama 2 was able to provide answers to simple questions. The main limitation was the hallucination effect. A hybrid approach, combining the strengths of multiple open-source LLMs (e.g., high accuracy, low hallucination tendency) shows great potential for developing robust, high-accuracy tools for emergency management and decision-support in the drinking water sector.
- **BeWaRE Benchmark, QMRA integration, and Decision-Support Tool:** The open-access BeWaRE benchmark testbed was created, integrating comprehensive knowledge on pathogen transport, chlorine decay, CRA degradation, stochastic water demands, and health impact assessment. This testbed utilized the QMRA framework by incorporating detailed exposure (via stochastic tap water consumption patterns) and dose-response models for wastewater contamination scenarios in both chlorinated and non-chlorinated DWDNs, offering a realistic representation of network dynamics and health outcomes. These efforts were translated into the PathoINVEST tool, a decision-support tool built on BeWaRE, enabling water utilities to manage

and model contamination events in real-time, find the source of contamination, and optimize mitigation strategies to improve public health.
- **Model-Based Emergency Response:** Traditional and model-based approaches were compared for managing a contamination event in a DWDN. This thesis demonstrated the applicability of PathoINVEST in a Case Study and showed that real-time modeling tools can improve contamination source identification (by reducing both the time required and the number of samples needed), optimize mitigation measures (via valve manipulation), and enhance decision-making during emergencies.

6.4 Implications for water utilities and practical applications

The research conducted and the methodologies that were developed in this thesis have implications for water utilities and offer practical applications that can improve emergency response and management of pathogen contamination events in DWDNs. During the PathoCERT project the AI-based methodology (Chapters 2 and 3) was translated into a web-based tool called PathoTHREAT[17][18]. Additionally, the BeWaRE benchmark testbed (Chapter 4) was incorporated into the PathoINVEST tool (Chapter 5).

6.4.1 Implications for water utilities

Improved decision-making

The use of an AI-based information extraction tool (PathoTHREAT) and a real-time modeling tool (PathoINVEST) enables water utilities to make faster and more informed decisions during emergencies. These tools significantly reduce the time needed to collect all relevant information about the emergency. Pathogen contamination events are (fortunately) rare, however, having key information on pathogen characteristics and earlier contamination events readily available is invaluable. Such key information includes the type of pathogen and their relevant symptoms, expected cases and hospitalizations, routes of exposure, and control options. It also improves the situational awareness of water utilities providing a better understanding of contamination spread, expected infection risk, and helping them to act quickly to mitigate risks by closing valves or isolating affected areas, which is crucial in minimizing the infection risk.

Operational efficiency

The incorporation of AI-based information extraction and modeling tools in the emergency response of water utilities significantly enhances their operational efficiency during contamination events.

- The AI-based information extraction tool eliminates the need for manual time-consuming literature reviews on similar historical contamination events, giving access to up-to-date information and saving water utilities valuable time during the critical early stages of contamination.
- Modeling tools such as PathoINVEST enables water utilities to:
    - Visualize the expected contamination propagation.
    - Focus on specific areas by guiding sampling locations and valve closures.
    - Reduce the number of valves required to close.
    - Reduce the number of samples required to find the contamination source.
    - Minimize the time needed to identify a contamination source in their network.
    - Minimize the time needed to decide on effective control options.

This approach can increase the overall efficiency, saving both time and costs during emergencies, reducing the health impact and number of people affected.

---

[17] https://github.com/Applied-Artificial-Intelligence-Eurecat/PathoTHREAT
[18] https://youtu.be/NTdF_aLsYL4?si=Czk5lC00aB31KKE1

Transferability

PathoTHREAT is applicable in any emergency situation that involves drinking water contamination. It provides a comprehensive overview tailored to effectively address a wide range of contamination events. PathoINVEST tool is designed to be adaptable since any DWDN with a suitable Geographical Information System and hydraulic modeling can support its use, regardless of network structure, or water system.

Integration into Standard Operating Procedures (SOPs)

Both PathoTHREAT and PathoINVEST can be smoothly integrated into the existing SOPs of water utilities, enabling them to adopt advanced decision-support systems without causing any disruptions to their current practices.

Sustainability

The functionalities and applicability of the PathoINVEST tool enable safe drinking water for the population, increasing trust in water supply and reducing reliance on bottled water. The rapid detection and isolation of contamination help minimize the affected area, leading to less water wasted during network flushing and disinfection. The tool can assist in minimizing the chemicals used for disinfection and streamlining decision-making to optimize resource allocation. In combination with efficient decision support, these features can contribute to less water loss and improved resilience against contamination events.

6.4.2 Practical applications

The above implications can be translated into practical applications in real-world contamination scenarios. Through the EU-funded PathoCERT project, three case studies of (hypothetical) contamination were conducted to showcase the usefulness and relevance of the developed tools during a contamination emergency.

PathoTHREAT

The web-based PathoTHREAT tool enables the user to import the extracted (AI-based) information in a database and using a User Interface to have comprehensive overview of pathogen characteristics and historical data of contamination events that are related to the current emergency, providing a rapid (high level) health and threat assessment. The tool was tested in all three case studies of the PathoCERT project, enabling relevant authorities to have a rapid assessment of their own contamination event and providing them with important information at the early stages of the emergency.

PathoINVEST[19]

The tool was applied in three European case studies (Spain, Cyprus, the Netherlands), each featuring distinct characteristics. In Spain and Cyprus, it was used to assist the response to earthquakes that led to sewage infiltration into the DWDN. In the Netherlands, it was employed to investigate suspected intentional contamination following customer complaints[20]. In each case study, emergency response teams comprising individuals from all relevant sectors (water utilities, civil protection, and health care) used PathoINVEST alongside their standard operating procedures.

Spain

In Spain, the water utility used the pathogen evolution forecasting feature to identify areas of concern within the network. Once they had a clear picture of the spreading, they used the PathoINVEST tool to identify sampling locations and verify the contamination. They also estimated the impact of contamination and proceeded with mitigation measures (valve manipulation) using the PathoINVEST feature to test different scenarios until they found the optimal valve closure combination. The responsible authorities noted that they could easily redeploy to new locations within such a short time

---

[19] Besides the features presented in Chapter 5 (sampling location suggestion, identify the contamination source, and optimal valve closure recommendation), the PathoINVEST tool includes another important feature, that is pathogen evolution forecasting.

[20] Part of the activities in this case study is presented in Chapter 5.

frame under proper communication. They also mentioned that it is a powerful tool that can be easily integrated into their standard operating procedures and in the hands of an expert, they can effectively mitigate any future contamination event in their network.

Cyprus

In Cyprus, after several locations in the DWN were considered as potential contamination sources, the PathoINVEST tool was used to suggest sampling locations to find the suspected source of contamination. Again, the feature of impact calculation was used once the source was identified to estimate the magnitude of the contamination after the earthquake. Finally, they were able to effectively restrain the contamination plume by strategically closing valves, as suggested by PathoINVEST. The responsible authorities acknowledged that their situational awareness and response had improved significantly compared to previous experiences, since they were able to immediately identify the source and prioritize effective mitigation measures.

Netherlands

In the Netherlands, customer complaints of water taste and odor were used as a starting point to identify the contamination source. This rapid identification enabled the water utility to respond faster and more efficiently. Their feedback was that in such situations (suspected intentional contamination), having a tool that models in real-time pathogen propagation and identifies the contamination source gave them a significant head start as it is crucial to act quickly. They also recognized the need for further training to effectively incorporate the tool into their standard operating procedures.

6.5 Perspectives and future research

6.5.1 Artificial Intelligence and LLMs for decision-making

The use of AI for information extraction has demonstrated significant potential for improving the situational awareness of responsible authorities during emergencies. The use of LLMs in QA tasks in a growing number of scientific fields (e.g., education, healthcare) is gradually becoming a necessity (He et al., 2023; Sajja et al., 2024; Yu et al., 2024). Being able to rapidly extract important information from scientific publications (asking a model using a prompt and receiving a response) instead of performing time-consuming literature reviews is considered extremely beneficial.

With respect to future research, translating AI-based methodologies for rapid information extraction into user-friendly tools that can be used from water utilities without requiring substantial training, represents a natural next step for real-world emergency management. Additionally, great attention should be paid to the responsible and domain-specific application of LLMs. In particular, their applicability to the domain of environmental microbiology requires careful investigation and necessitates human supervision (i.e., human-in-the-loop). Furthermore, there is an imperative need for improved documentation of contamination incidents. Currently, many events are either underreported or not documented as contamination events or outbreaks. And even when incidents are reported, the information is often not properly structured, making the application of AI for information extraction a difficult task. If incident reporting were more structured and made accessible in the public domain, AI could play an even greater role in harnessing and analyzing this information, similar to what is already happening in the biomedical domain. Although the rapid progression of AI and specifically LLMs show great potential and is already impacting our lives, there is also a risk that society, including experts, may struggle to fully harness their potential. An example is the hallucination effect as mentioned in Chapter 3, where LLMs generate plausible yet wrong answers, highlighting the importance of addressing these challenges to ensure a responsible and efficient use of AI. It is important to emphasize that AI is not meant to replace human decision-making but to support and enhance it. Finally, the digitalization of the water sector is already underway, with AI being an integral component of this transformation. Integrating AI into the standard operating procedures of water utilities seems not only a logical but a necessary next step.

6.5.2 Modeling tools for decision-making

The literature review has shown that while a lot of progress has been made in modeling the various reactions that take place in a DWDN, existing efforts have predominantly focused on specific aspects of hydraulic and quality dynamics modeling, often relying on assumptions and simplifications. This thesis demonstrated that it is feasible to develop a holistic representation of all the reactions occurring within a network, by utilizing state-of-the-art methodologies and developing a new benchmark testbed. The proposed approach integrates key factors -chlorine decay, reaction kinetics, CRA degradation, and stochastic water consumption patterns- into a single testbed, offering a more realistic and comprehensive alternative to traditional models that often oversimplify or isolate these dynamics. This integration enables users (water utilities) that have a hydraulic network to realistically simulate any contamination event and assess its impact (through the framework of QMRA). This thesis also showed that the use of real-time modeling is beneficial, as delayed responses can hinder any meaningful mitigation measures during an emergency in the DWDN.

The ability to realistically simulate a contamination event in the DWDN, while having information in real-time about the pathogen propagation, the potential infection risk, as well as a proposal for sampling and the optimal closure of valves for mitigation is invaluable for a water utility. The current approach regarding emergency management in the DWDN is highly dependent on experience and human intuition rather than what is actually happening in the network as the contamination event unfolds (use of real-time data). Chapter 5 underscored the added value of integrating modeling tools to enhance the current practices of water utilities, shifting decision-making from assumption-driven to data-driven approaches.

The strengths of modeling tools have been clearly demonstrated in this thesis, but there are also weaknesses that need to be addressed in the future. For instance, future research should focus on understanding the dynamics within a DWDN when chlorine-reducing agents found in wastewater (e.g., organic/inorganic compounds and pathogens) "compete'' for chlorine. Key questions include: What are the actual disinfection kinetics when wastewater enters a chlorinated network? Which reactions occur first, and which follow? Do these reactions happen simultaneously, or sequentially? Is it valid to assume that individual disinfection kinetics can simply be combined? Additionally, current hydraulic modeling relies on the perfect mixing assumption within pipes. Is this assumption accurate, or does it oversimplify the real behaviour within the network? And does this influence the accuracy of the predicted health impact or is the effect of the assumption negligible? Addressing these questions using experimental designs will be crucial for advancing the reliability and applicability of these modeling tools in real-world scenarios.

A better understanding of contamination dynamics also requires access to real data on chlorine-reducing agents in wastewater during contamination events. To address this gap, future research should focus on pilot case studies to provide experimental data, enabling the calibration of the BeWaRE testbed with real DWDN data. These experimental studies could explore various hydraulic and water quality conditions, emphasizing the interactions between chlorine and chlorine-reducing agents across different phases (fast and slow). Additionally, more attention should be paid to wall chlorine decay dynamics, as these vary significantly between networks depending on the pipe material, age, and condition.

Regarding the health impact, as discussed in Chapter 4, the infection risk estimation needs to be more comprehensive by exploring a cumulative infection risk from multiple pathogens during a wastewater contamination event. To be able to do this properly, more data on the concentration of infectious pathogens in different DWDN contamination sources is needed. Additionally, for a holistic QMRA implementation, the exposure via showering and brushing teeth should be taken into account.

To further improve the accurate outputs of the BeWaRE testbed, different pathogens and dose-response models could also be explored as well to better capture the potential infectivity of a wastewater contamination event. The difference between the probability of infection and the probability of illness should be explored since these two are distinct concepts in dose-response modeling, each representing different outcomes after pathogen exposure. While the first represents the likelihood that an individual will be infected after pathogen exposure (and after a specific dose), the latter refers to the likelihood that an infected individual will become ill after exposure. Infection is a prerequisite for illness, but infection does not guarantee illness. This means that the probability of illness is lower than the probability of infection. This distinction could help refine risk management strategies by enabling responses to be tailored to vulnerable groups of the population (e.g., prioritize mitigation measures such as valve closing or re-chlorination in areas with hospitals). This differentiation could also have implications to the refinement of dose-response models, separating infection and illness probabilities for critical pathogens and vulnerable groups. Subsequently, the inherent uncertainty of transitioning from infection to illness could be incorporated and taken into account when combining QMRA with hydraulic modeling. Additionally, the Disability-Adjusted Life Year (DALY) metric could also be explored and integrated into future studies combining modeling and health risk assessment. The DALY framework combines the years of life lost due to premature mortality and the years of living a life with disability due to a disease, providing a quantitative metric of health burden associated with contamination. Incorporating DALYs when modeling health risk could enhance the dissemination of such critical information to water utilities. It could also allow for better comparison of health risks between different pathogens, but also assess and communicate the long-term effects of a contamination, providing a better understanding of the trade-offs between different mitigation strategies.

A shift in the status quo of DWDN management (both for routine operations and emergencies) is essential to address limitations in current practices, such as delayed detection of issues and reactive responses. The digitalization of the water sector offers a powerful means to achieve this transformation. Digital Twins, for example, can monitor water quality in real-time, identifying critical locations and vulnerabilities within a network, thus enabling a more proactive and effective management (Karmous-Edwards et al., 2019). Increasing the detection frequency of contamination events and outbreaks in the

DWDN as they unfold and not after a critical point, as discussed in Figure 1.1 (Chapter 1), is extremely important. Literature review showed that only 7% of contamination cases are detected through customer complaints, another 7% via water quality analysis, and 86% through unusual increases in hospitalizations. Future emergency water management should consider deploying online sensors that can detect microbial contaminants in real time throughout the DWDN, to enable rapid responses (Raich, 2013; Gunnarsdottir et al., 2020; Canciu et al., 2022). Building on this rationale, findings in these BTO reports (Blokker et al., 2017a, 2017b, 2018a, 2018b) highlight the imperative need for developing automatic, online microbial sensors in the DWDN. While this research has shown that online microbial sensors can be very beneficial in exceptional situations (e.g., wastewater contamination in the DWDN), they can also be beneficial in routine operations (e.g., main repairs) since they can reduce the contamination detection time (e.g., 1-4 hours post-repair compared to 12-24 hours). Although online sensors are considerably more expensive, they can significantly improve detection probabilities (from 5% with conventional sampling to 65% when using online sensors). Additionally, under specific conditions these sensors can lead to a 94% reduction in contaminated areas and a 50% reduction in the contamination duration. Therefore, future research should prioritize evaluating whether the benefits of deploying online sensors outweigh their additional costs, in terms of improved public health protection and operational efficiency.

The development of Digital Twins is beneficial for water utilities since they enable real-time network hydraulics monitoring, scenario planning, and modeling (Karmous-Edwards et al., 2019). Ensuring these systems reflect the accurate status of the network characteristics (having an up-to-date network), including the state of valves, is essential; modeling tools are only as reliable as their input data. Ultimately, reaching to a point where experts have full situational awareness over their critical infrastructure (DWDN) -including interconnections and interdependencies with transportation, power, communication and other essential systems- will transform water utility operations and emergency responses for the better.

References

Blokker, E.J.M., Mesman, G.A.M. Responsstrategie na meting E.coli door sensoren. Response Strategy After E.coli Detection by Sensors. 2017a, 20 pages, KWR, Nieuwegein, Netherlands

Blokker, E.J.M., Smeets, P.W.M.H. Toegevoegde waarde online E. coli sensor in het distributienet. Added Value of Online E. coli Sensors in the Network. 2017b, 44 pages. KWR, Nieuwegein, Netherlands.

Blokker, E.J.M., Smeets, P.W.M.H., Fujita, Y. Implementatieplan snel meten na werkzaamheden. Implementation Plan for Rapid Sampling After Maintenance. 2018a, 18 pages. KWR, Nieuwegein, Netherlands.

Blokker, E.J.M., Smeets, P.W.M.H., Hessels, L. Implementatie automatische snelle detectie van fecale verontreiniging in het distributienet. Implementation of Automatic Rapid Detection of Fecal Contamination in the Network. 2018B, 29 pages. KWR, Nieuwegein, Netherlands.

Gunnarsdottir, M. J., Gardarsson, S. M., Figueras, M. J., Puigdomènech, C., Juárez, R., Saucedo, G., ... & Hunter, P. (2020). Water safety plan enhancements with improved drinking water quality detection techniques. *Science of the total environment*, *698*, 134185.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E. (2023). A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Karmous-Edwards, G., Conejos, P., Mahinthakumar, K., Braman, S., Vicat-Blanc, P., & Barba, J. (2019). Foundations for building a digital twin for water utilities. *Smart Water Report–Navigating the smart water journey: From Leadership To Results, Water Online, SWAN*, 9-20.

Raich, J. (2013). Review of sensors to monitor water quality. *European reference network for critical infrastructure protection (ERNCIP) project*.

Sajja, R., Sermet, Y., Cikmaz, M., Cwiertny, D., & Demir, I. (2024). Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education. *Information*, *15*(10), 596.

Yu, H., Yu, C., Wang, Z., Zou, D., & Qin, H. (2024). Enhancing healthcare through large language models: A study on medical question answering. *arXiv preprint arXiv:2408.04138*.

List of publications and conferences

Publications

Paraskevopoulos, S., Smeets, P., Tian, X., & Medema, G. (2022). Using Artificial Intelligence to extract information on pathogen characteristics from scientific publications. *International Journal of Hygiene and Environmental Health*, *245*, 114018.

Paraskevopoulos, S., Vrachimis, S., Kyriakou, M., Pavlou, P., Kouzapas, D., Milis, G., Smeets, P., Eliades, D., Medema, G., Polycarpou, M., & Panayiotou, C. (2022). PathoINVEST: Pathogen contamination investigations during emergencies. In 2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Industry. Universitat Politecnica de Valencia.

Paraskevopoulos, S., Vrachimis, S., Kyriakou, M., Eliades, D. G., Smeets, P., Polycarpou, M., & Medema, G. (2024). Modeling the health impact of wastewater contamination events in drinking water networks. *Journal of Cleaner Production*, 143997.

Paraskevopoulos, S., Vrachimis, S. G., Kyriakou, M. S., Blokker, M., Smeets, P., Eliades, D. G., ... & Medema, G. (2024). An Innovative Model-Based Methodology for Rapid Response to Drinking Water Contamination Events. *Engineering Proceedings*, *69*(1), 45.

Paraskevopoulos, S., Vrachimis, S., Kyriakou, M., Eliades, D., Smeets, P., Blokker, M., ... & Medema, G. Addressing Drinking Water Contamination: A Case Study Comparing Traditional with Model-Based Approaches. *Available at SSRN 4877088*. **(Preprint)**

Conferences

“Artificial Intelligence and emergencies: An automated approach to extract information from the literature to tackle pathogen contamination events”, Amsterdam International Water Week, Amsterdam, the Netherlands, 5 November 2021.

“PathoINVEST: Pathogen Contamination Investigations during Emergencies”, 2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2022), Valencia, Spain, 18–22 July 2022.

“Modeling impact and mitigation of drinking water network microbial contamination events”, Health and Environment 2023, Cannes, France, 26-28 June, 2023

“An Innovative Model-Based Methodology for Rapid Response to Drinking Water Contamination Events”, 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

About the author

Sotirios Paraskevopoulos was born in Larisa, Greece. His deep appreciation for nature led him to enroll in the Department of Forestry and Natural Environment at Aristotle University of Thessaloniki, Greece, in the summer of 2007. During his studies, he developed a strong interest in environmental systems and management, ultimately earning his degree in 2014. Shortly after, he pursued his first master’s degree at the University of Thessaly, Greece, in “Spatial Analysis and Environmental Management.” For his thesis, titled “Spatial Planning of Constructed Wetlands using Multi-Criteria Analysis: A Case Study of the Municipality of Kileler, Greece,” he explored innovative methods for optimizing constructed wetlands placement to enhance environmental management.

Driven by a growing passion for water management, Sotirios embarked on a life-changing journey in 2017 to pursue his second master’s degree at Utrecht University, the Netherlands, in "Water Science and Management". His thesis, “Quantifying the Log Reduction of Pathogenic Microorganisms by Constructed Wetlands as a Basis for QMRA of Water Reuse Applications,” was conducted during an internship at KWR Water Research Institute. This work marked a significant step in his specialization, focusing on integrating Quantitative Microbial Risk Assessment (QMRA) into water reuse applications.

In 2020, while continuing his work as a researcher at KWR, Sotirios began his PhD journey at Delft University of Technology, the Netherlands, in the Department of Water Management, Faculty of Sanitary Engineering. His doctoral research was supported by the Horizon 2020 PathoCERT project and focused on modeling waterborne pathogens in drinking water distribution networks (DWDN). Sotirios developed expertise in modeling pathogen inactivation kinetics, stochastic water demands, chlorine decay, and QMRA. His research also explored innovative mitigation strategies, including the use of evolutionary algorithms to optimize valve manipulation during contamination events.

At the beginning of 2022, Sotirios expanded his expertise further as a visiting researcher at the KIOS Research and Innovation Center of Excellence, University of Cyprus. There, he collaborated with leading experts to enhance his work on modeling pathogen behavior in DWDN. His time at KIOS contributed significantly to refining his models and understanding the complexities of pathogen transport and inactivation in chlorinated systems.

Beyond his contributions to modeling and risk assessment, Sotirios actively worked on enhancing rapid response capabilities for water contamination events. His work involved developing Artificial Intelligence-based applications for extracting critical information about waterborne pathogens, aiding in real-time threat and risk assessment during emergencies.

Sotirios’s research has practical implications for water utilities and policy development, addressing global challenges in drinking water safety and emergency response.

Outside academia, Sotirios enjoys spending time in nature, a passion that has remained constant throughout his life. He continues to seek new opportunities to combine his expertise in water management with cutting-edge technologies.

Acknowledgements

The journey to a PhD is considered (arguably so) a lonely journey. The candidate must put in the hard work, face struggles, pull all-nighters, work on weekends, and meticulously address reviewer's comments. However, what happens behind the scenes paints a very different picture. The famous saying "it takes teamwork to make the dream work" perfectly captures the essence of PhD research. My journey was supported by a team of individuals who contributed in various ways through cooperation, guidance, constructive criticism, and emotional support. The completion of this PhD would not have been possible without them, and I would like to express my gratitude by acknowledging their contributions.

I would like to thank my **KWR colleagues from the Microbial Water Quality and Health team**. You provided a great working environment for a young researcher, which enabled me to maintain (as much as possible) a great work-life balance. It was always nice to share frequent updates on my progress -and struggles- over the past four years.

A special thank you to **Dr. Lydia Vamvakeridou** for persuading me to accept the PhD position by highlighting the endless possibilities that awaited me once this journey was over. You were absolutely right! **Lydia, Konstantine, Alex, Julian, Sid, Tavishi** and the rest of the Watershare Associates team, thank you so much for the fun conversations by the coffee machine, for sharing PhD struggles, and for the fun nights in Utrecht and conferences.

I would also like to thank **Dr. ir. Mirjam Blokker**, my weekly supervisor, for sharing her knowledge and always being there for me to answer questions about water demands and hydraulic modeling.

A big thank you goes to **Prof. dr. ir. Jan Peter van der Hoek** and **Prof. dr. Zoran Kapelan** for their valuable contributions and constructive feedback on my thesis. Your insights have been valuable during this research journey.

Finally, I would like to thank my buddies from the Netherlands whom I met during my master's studies. **Oscar, Bart, Jackie boy**, thank you for the wonderful years in Utrecht. Cheers!

A special chapter of this PhD (and my life) is my Cypriot family at the KIOS Research and Innovation Center of Excellence, University of Cyprus. Without their support, this journey would have been totally different (and most likely much slower).

Demetri, thank you for welcoming me with open arms while I was still trying to figure out where my research is heading. Your willingness to share your knowledge, wisdom, and brilliant ideas with me paved a path that I could never have imagined. I believe I have told you in person, you are an entrepreneur, an influential researcher that I am grateful to have met.

Stelio, my dear friend, a big thank you for helping me find my way, for calming me down whenever I vented about reviewer's comments, for sharing your experiences on PhD life, and for teaching me so much about graph theory, hydraulic modeling, as well as providing an engineer's perspective to the problems I was trying to solve. Your support has been invaluable.

Marie, thank you for the countless hours we spent together in front of our laptops, writing code, creating innovative applications, and solving unexpected problems with "hakies"! I know I sometimes asked too much of you, especially when requesting last-minute fixes for the PathoINVEST tool, but you were always there to help. Thank you for your patience and constant willingness to help!

Nothing would have been possible without the support of my promotor and colleague Prof. dr. Gertjan Medema, and my colleague, supervisor and mentor Dr. ir. Patrick Smeets.

Gertjan, I can now confidently say that I couldn't have asked for a better promotor and supervisor. Thank you for your guidance, wisdom, and knowledge sharing, but most importantly, for letting me create my own path. You gave me space to grow and provided what I now consider the best advice: "Suck it up and move on". This advice goes beyond rejected publications and has become a valuable life lesson. I know we had a rough start, but as soon as I realized that I was responsible for my own research, everything changed. I learned how to listen, think critically, and take ownership of my work. From that point onward, our communication and collaboration reached an entirely new level, for which I am truly grateful.

Patrick, I consider myself lucky to have met you. It all started back in 2017, when I arrived late to your class at the University. Bursting into the room mid-lecture, scrambling to find a seat, and unintentionally disrupting the flow, I distinctly remember your playful comment, “Look who decided to join us!” From that moment on, a unique bond was formed. Little did I know that this encounter would lead to years of inspiration, mentorship, collaboration, and friendship. Thank you for always being there for me during my internship, during this PhD journey with endless discussions, and during my personal struggles both in and beyond academia. Thank you for everything you’ve taught me -not just about science, but about life.

Words cannot express the depth of gratitude I feel for my family because, without them, I wouldn’t be where I am today. **Μαμά, Μπαμπά, Γιώργο, Αλέξανδρε**, σας ευχαριστώ για όλα. Ευχαριστώ που στηρίξατε από την αρχή το (τότε) νέο μου ξεκίνημα στην Ολλανδία. Ευχαριστώ που ήσασταν πάντα κοντά μου, ακόμα και αν μας χώριζαν χιλιάδες χιλιόμετρα. Έχετε μεγάλο μερίδιο σε όλο αυτό. Ευχαριστώ που πιστέψατε σε μένα. Τα κατάφερα! Σας αγαπώ!

And last but certainly not least to my wife **Chrysi** -my partner, my family, my friend, and my constant source of strength- my eternal gratitude goes to you. Thank you for being the steady rock I needed to see this through, for your patience as I traveled back and forth between Greece and the Netherlands, for using your artistic talent to help me polish the figures in my publications, and for always lifting me up when my publications were rejected! Your unwavering support, love, and encouragement were exactly what I needed to complete this journey. Αγαπημένη μου γυναίκα, σε ευχαριστώ για όλα! Σε ευχαριστώ που πιστεύεις σε μένα, γιατί αυτό με εμπνέει να προσπαθώ να γίνομαι συνεχώς καλύτερος. Σε αγαπώ!