P.PORTO

ESCOLA
SUPERIOR
DE SAÚDE
POLITÉCNICO
DO PORTO

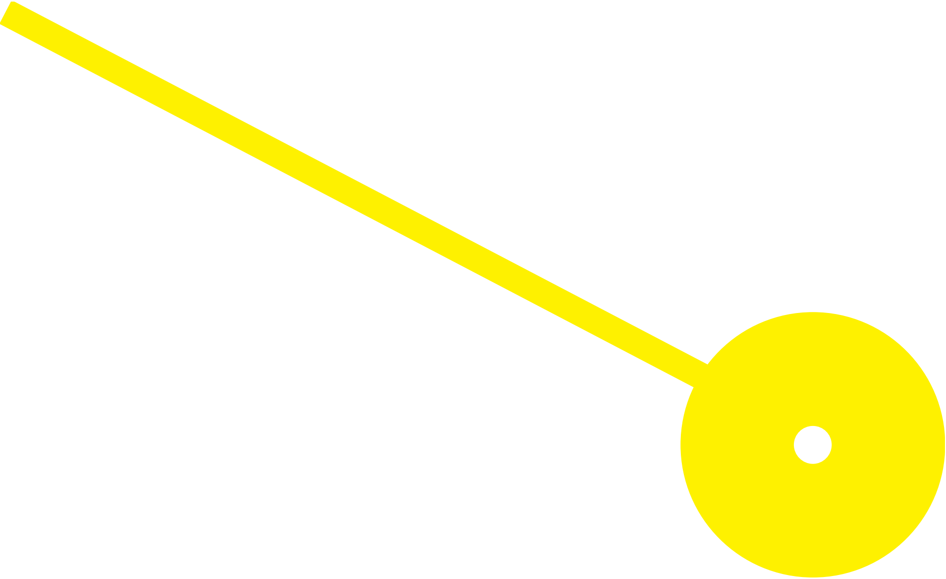
M

MESTRADO
HIGIENE E SEGURANÇA NAS ORGANIZAÇÕES

Predictive models for evaluation of *Legionella spp.*: a systematic review

Bárbara Beatriz Andrade Carneiro

10/**2024**



P.PORTO

ESCOLA
SUPERIOR
DE SAÚDE

Predictive models for evaluation of *Legionella spp.*: a systematic review

Autor

Bárbara Beatriz Andrade Carneiro

Orientador

Professora Doutora Maria Manuela Vieira da Silva, Escola Superior de Saúde do Instituto Politécnico do Porto

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em **Higiene e Segurança nas Organizações** pela Escola Superior de Saúde do Instituto Politécnico do Porto.

Agradecimentos

Um muito obrigada a todos que participaram, direta ou indiretamente no desenvolvimento deste trabalho, enriquecendo o meu processo de aprendizagem. Aos meus pais e irmão por toda a força que me deram e à minha colega Inês pela sua disponibilidade e apoio em todos os momentos.

Por fim, um especial obrigada à Professora Doutora Maria Manuela Vieira da Silva por toda a orientação e pelo interesse que me incutiu neste trabalho.

Resumo

Introdução: *Legionella spp.* é uma importante bactéria transmitida pela água, estando na origem na doença dos legionários. Os modelos preditivos são cruciais para avaliar os riscos de contaminação e prevenir surtos. Esta revisão sistemática tem como objetivo identificar os diversos modelos preditivos utilizados direta ou indiretamente para a avaliação da presença de *Legionella*, assim como avaliar a fiabilidade dos modelos preditivos.

Métodos: Seguindo as diretrizes do PRISMA, esta revisão incluiu estudos de 2014 a 2024, obtidos nas bases de dados PubMed, ScienceDirect e Web of Science, focando-se em modelos preditivos de *Legionella*. A metodologia SPIDER foi utilizada para definir os critérios de elegibilidade e a Checklist de Avaliação Crítica da JBI foi aplicada para avaliar a qualidade metodológica dos estudos.

Resultados: Foram incluídos dezassete estudos que utilizaram modelos de 'machine learning' e modelos estatísticos tradicionais. As variáveis de destaque que influenciam o crescimento de *Legionella* spp. foram a temperatura da água, os níveis de cloro, a presença de biofilme e condições ambientais como a humidade. Os modelos de 'machine learning' demonstraram maior precisão, mas apresentaram complexidade na interpretação, enquanto os modelos tradicionais ofereciam mais transparência, mas careciam de precisão em ambientes dinâmicos.

Discussão: A revisão constatou que os modelos preditivos que incorporam variáveis ambientais, biológicas e infraestruturais oferecem um potencial significativo para prevenir surtos de *Legionella*. Os modelos de 'machine learning' demonstraram maior precisão, mas a sua complexidade apresentou desafios para a implementação prática. Os modelos tradicionais, embora menos complexos, oferecem maior transparência, mas são limitados em precisão. Além disso, fatores relacionados com as alterações climáticas, como o aumento das temperaturas e da precipitação, provavelmente aumentarão os riscos de contaminação por *Legionella spp.*, sublinhando a necessidade de modelos preditivos mais adaptáveis.

Conclusão: Embora nenhum modelo isolado seja universalmente aplicável, a integração de abordagens de 'machine learning' e métodos estatísticos mostra-se promissora. Esforços futuros devem focar-se no refinamento dos modelos para melhorar a sua generalização e adaptação às condições ambientais em mudança.

Palavras-chave: *Legionella spp.*; *Legionella pneumophila*; modelos preditivos; alterações climáticas; saúde pública.

Abstract

Introduction: *Legionella spp.* is a significant waterborne bacterium, responsible for Legionnaires' disease. Predictive models are crucial for assessing contamination risks and preventing outbreaks. This systematic review aims to identify the various predictive models used directly or indirectly to evaluate the presence of *Legionella*, as well as assess the reliability of these predictive models.

Methods: Following PRISMA guidelines, this review included studies from 2014 to 2024, retrieved from PubMed, ScienceDirect, and Web of Science, focusing on predictive models for *Legionella*. The SPIDER methodology was used to define eligibility criteria, and the JBI Critical Appraisal Checklist was applied to evaluate the methodological quality of the studies.

Results: Seventeen studies were included, utilizing machine learning models and traditional statistical models. Key variables influencing the growth of *Legionella spp.* were water temperature, chlorine levels, biofilm presence, and environmental conditions such as humidity. Machine learning models demonstrated greater accuracy but posed challenges in interpretation, while traditional models offered more transparency but lacked precision in dynamic environments.

Discussion: The review found that predictive models incorporating environmental, biological, and infrastructural variables offer significant potential in preventing outbreaks of *Legionella*. Machine learning models demonstrated higher accuracy, but their complexity posed challenges for practical implementation. Traditional models, while less complex, offered greater transparency but were limited in precision. Additionally, climate change-related factors, such as rising temperatures and increased rainfall, are likely to increase contamination risks from *Legionella spp.,* underscoring the need for more adaptable predictive models.

Conclusion: While no single model is universally applicable, the integration of machine learning approaches with statistical methods appears promising. Future efforts should focus on refining these models to improve their generalization and adaptation to changing environmental conditions.

Keywords: *Legionella spp.*; *Legionella pneumophila*; predictive models; climate change; public health.

Table of Contents

List of Abbreviations

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

JBI – Joanna Briggs Institute

SPIDER – Sample, Phenomenon of Interest, Design, Evaluation, Research Type

CT – Cooling Tower

CFU – Colony-Forming Units

ROC – Receiver Operating Characteristic

CDC – Centers for Disease Control and Prevention

EPA – Environmental Protection Agency

WHO – World Health Organization

QUADAS – Quality Assessment of Diagnostic Accuracy Studies

QMRA – Quantitative Microbial Risk Assessment

EPANET-MSX – EPANET Multi-Species Extension

BAC – Baltimore Aircoil Company

GBHM – Great Britain Human Movement database

AI – Artificial Intelligence

MPN – Most Probable Number

DALY – Disability-Adjusted Life Years

UN – United Nations

List of Tables

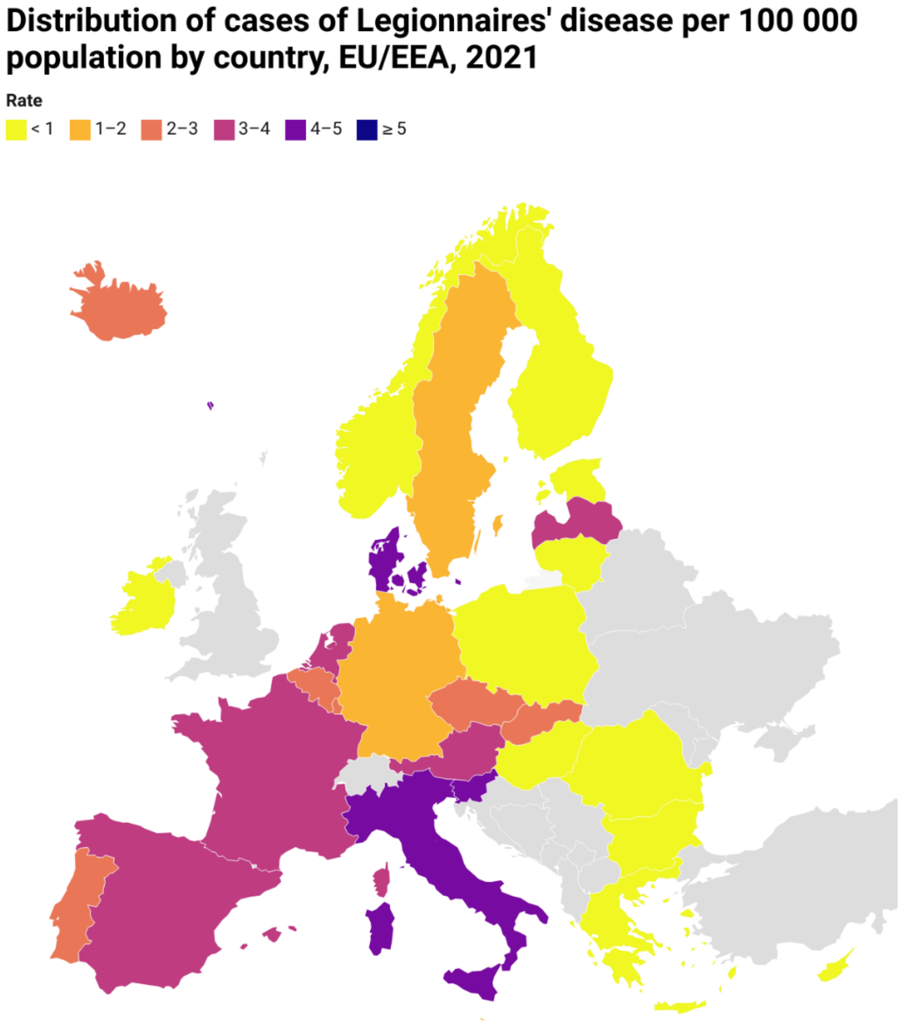
List of Figures

1. Introduction

In a constantly evolving world, water safety is of the utmost importance. It has become an increasingly critical issue, particularly as the population grows and urbanizes. The world's population is expected to continue growing over the coming five decades, reaching a peak of around 10.3 billion people in the mid-2080s (UN, 2024). With more people relying on shared water resources, contamination, and waterborne diseases have heightened, necessitating stringent safety measures and regulations. Water is vital for sustaining life, but when it becomes contaminated, it accounts for 4% of human deaths and 5% of global health issues (Sharma, et al., 2023). The United Nations Sustainable Development Goals (SDGs) include Goal No. 6: Clean water and sanitation. It aims to ensure the availability and sustainable management of water and sanitation for all by 2030 (UN, 2023). This goal is critical, as water and sanitation are fundamental to human health, environmental sustainability, and economic development.

Contaminated water can act as a passive carrier for infections, causing illnesses known as waterborne diseases when consumed. These diseases may be triggered by viruses, bacteria, or parasites. Water contaminants pose significant health risks, particularly for infants, young children, pregnant women, and individuals with compromised immune systems (Sharma, et al., 2023). Over the past decade, the incidence of waterborne diseases has increased. This rise is linked to various factors, including demographic changes such as an aging population, and the increasing population density in urban areas with outdated centralized water systems (NASEM, 2020), climate change, and improved reporting and detection methods.

For instance, Legionnaires' disease, caused by the *Legionella* bacterium, has become one of the most reported waterborne diseases in the European Union and the European Economic Area (Kanarek, Bogiel, & Breza-Boruta, 2022). In the Annual Epidemiological Report, from the European Centre for Disease Prevention and Control (ECDC) published in July 2023, 29 countries reported 10 723 cases of Legionnaires' disease, of which 10 004 (93%) were classified as confirmed (ECDC, 2023). The number of notifications per 100 000 population increased to 2.4, which was higher than any precedent year under surveillance. Four countries, Italy, France, Spain, and Germany, continued to account for the majority of notified cases (75%), although their combined populations represented approximately 50% of the EU/EEA population. Out of 8 054 cases with known outcome, 704 (9%) were reported to have a fatal outcome.



Source: European Centre for Disease Prevention and Control, 2023. • Created with Datawrapper

Figure 1 – Distribution of cases of Legionnaires' disease per 100 000 population by country in 2021 (ECDC, 2023)

Advances in technology and infrastructure have allowed for improved water treatment and monitoring systems, but these must be continually updated to keep pace with new contaminants and climate change impacts. By prioritizing the prediction of *Legionella* outbreaks, public health officials can implement more effective prevention and control strategies, ultimately reducing the incidence and impact of Legionnaires' disease.

Additionally, predictive models help allocate resources more efficiently, targeting prevention efforts where they are most needed, and reducing healthcare costs associated with treating

Legionnaires' disease. They also assist in compliance with regulatory standards, ensuring water systems are maintained and monitored effectively to safeguard public health.

The objective of this systematic review is to find out if any of the existing predictive models is reliable for prevent *Legionella* infections. The specific objectives are: (1) identify the existing predictive models (2) interpret each model (3) describe its vantage points (4) recognize its limitations (5) identify which variables are important for the proliferation of *Legionella spp.* (6) explain the importance of predicting the infection.

2. Bibliographic Review

2.1. *Legionella spp. and Legionella pneumophila*

Legionella spp. describes a gram-negative bacillus bacterium (Mraz & Weir, 2018). *Legionella* was first identified as a cause of human disease in 1976 following an outbreak of pneumonia among American Legion members at a conference held at the Bellevue-Stratford Hotel in Philadelphia. Among the nearly 2,000 people present, 182 individuals developed the disease, leading to 29 deaths (NASEM, 2020).

The recognized species and serogroups of the genus *Legionella* have been steadily rising since the Philadelphia outbreak in 1976. Currently, there are over 60 identified species and 80 serogroups (Miyashita, et al., 2020). Most reported cases are caused by *Legionella pneumophila* and most frequently serogroup 1. In people with weakened immune systems, species other than *L. pneumophila* that are frequently isolated include *L. micdadei*, *L. bozemanii*, *L. dumoffi* and *L. longbeachae* (NASEM, 2020).

Legionella spp. demonstrate substantial pleomorphism, adopting rod-shaped, coccoid, and filamentous forms depending on environmental factors like temperature, nutrient availability, presence of metabolites, and their habitat. Although these aerobic bacteria are widely distributed, they do not form spores (Kanarek, Bogiel, & Breza-Boruta, 2022). Throughout its life cycle, *Legionella* alternates between infectious and replicative forms, as well as more resilient, dormant cell forms (NASEM, 2020).

Table 1 – Microbiology of Legionella spp. Adapted from *(Mraz & Weir, 2018).*

Habitat	Widespread in nature, common to fresh surface water, thermal water, and drinking water environments and has been found regularly in soil samples.
Ideal temperature	Thrive in temperatures 20–50º C with little to no growth over 50º C but have been isolated from hot water with temperatures as high as 66 ºC. When temperatures reach 70 ºC, the bacteria are killed almost instantly.
Tolerances	Shown to be tolerant to sodium chloride solutions up to 1.5% laboratory experiments and isolated in amoeba co-cultures in natural environments of up to 3% sodium chloride.
pH	Ideal at 6.0–8.0.
Nutrients	Is tolerant of low-nutrient environments but does require iron to survive and thrive. The bacteria can enter a viable but non-culturable (VBNC), state in which it will not replicate, but can be resuscitated in more friendly conditions.

2.2. Symbiotic relationships

L. pneumophila tends to struggle when existing as solitary, free-living cells in natural environments. Rather, it thrives most effectively within amoebae and other free-living protozoa that reside in biofilms. This bacterium's growth requirements are well-suited to a parasitic lifestyle (NASEM, 2020).

Legionella that have inhabited protozoan hosts become more infectious to mammalian macrophages. This heightened infectivity is partly due to horizontal gene transfer between the bacteria and the host cell. (Mraz & Weir, 2018).

A biofilm consists of microorganisms enclosed within a self-generated, hydrated gel matrix that attaches to wet soil, sediments, and other solid surfaces, gathering both organic and inorganic materials. Biofilms commonly form on any damp surfaces, including human-made structures like pipes, tanks, filters, and gaskets — basically, anything that comes into contact with water. The material of the pipe can greatly influence the composition and behavior of the microbial community within the biofilm. Furthermore, both the type of surface and the temperature affect the complex interactions between *Legionella*, its amoebae hosts, and other organisms within the biofilm community (NASEM, 2020).

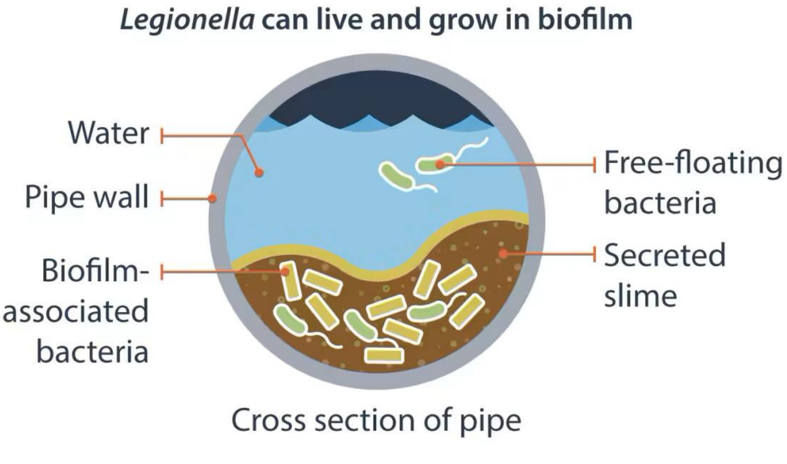


Figure 2 – Legionella growth in biofilm. From (CDC, 2024).

2.3. Natural and artificial reservoirs

Legionella spp. is autochthonous to natural freshwater environments and water systems, including lakes, rivers, damp soils, and decomposed plant matter (Kanarek, Bogiel, & Breza-Boruta, 2022).

Legionella spp. have been found on every continent (NASEM, 2020). An investigation into the microbial community of glacial lakes close to the Syowa research station in Antarctica also detected the presence of *Legionella spp*. in this environment. The research revealed that different *Legionella spp.* was extensively distributed in the Antarctic environment and were phylogenetically distinct from previously described species. This suggests that there is still many undiscovered *Legionella spp.* in Antarctica, indicating that the genus likely includes a broader range of species in cold environments than what is currently recognized (Shimada, et al., 2021).

Nevertheless, the uncontrolled proliferation of pathogenic *legionellae* in artificial water systems usually results in human exposure and illness. Areas where biofilms form and the risk of aerosolization exists can serve as potential sources of *Legionella* infection. (NASEM, 2020). Locations such as contaminated hospital bathrooms, grocery store misting machines, sprinklers, HVAC systems, humidifiers, large warm water systems, whirlpools, swimming pools, and dental office equipment (Mraz & Weir, 2018).

Cooling towers in buildings and industrial environments, such as power plants and factories, have been associated with multiple *Legionella* infection outbreaks (NASEM, 2020). Cooling towers, built to reduce water temperature in industrial systems, can generate bioaerosols that contain *Legionella* bacteria. Contaminated towers are capable of spreading these airborne particles across several kilometers (Kanarek, Bogiel, & Breza-Boruta, 2022).

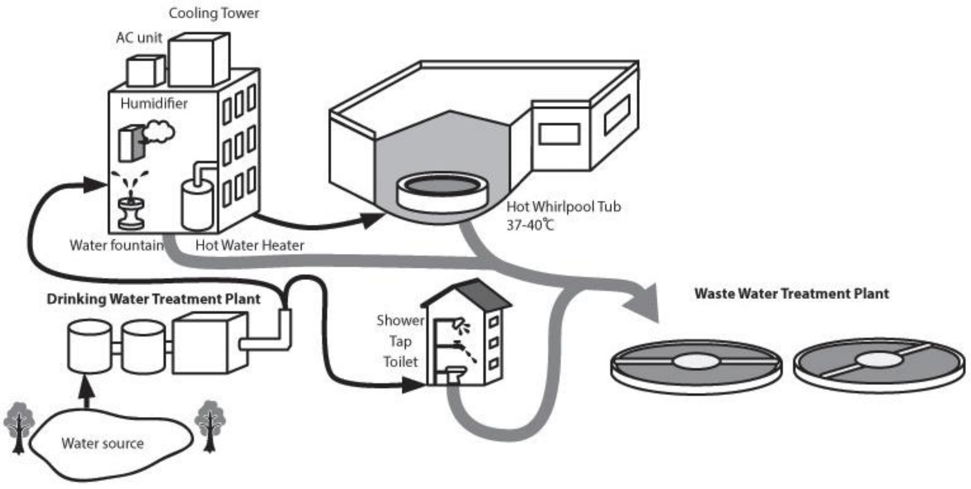


Figure 3 – Built locations where Legionella growth can occur. NOTE: Thinner black arrows indicate water pathways to premise plumbing and thicker grey arrows indicate wastewater. From (NASEM, 2020).

2.4. *Legionella* infection

Individuals infected with *Legionella* can develop pneumonia, referred to as Legionnaires' disease, or a milder flu-like illness called Pontiac fever; together, these conditions are known as legionellosis (NASEM, 2020). Legionellosis can only be spread when the bacterium is aerosolized and inhaled (Mraz & Weir, 2018). Aerosols are small liquid droplets, typically under 100 µm in diameter, formed by fluid turbulence. However, only droplets smaller than 10 µm can reach deep into the human lungs. These droplets can carry suspended substances, such as bacteria and protozoa. Once the bacteria enter the respiratory system, they replicate in pulmonary macrophages and monocytes. The incubation period is generally believed to range between two to twelve days but can be longer, especially in immunocompromised individuals. (NASEM, 2020).

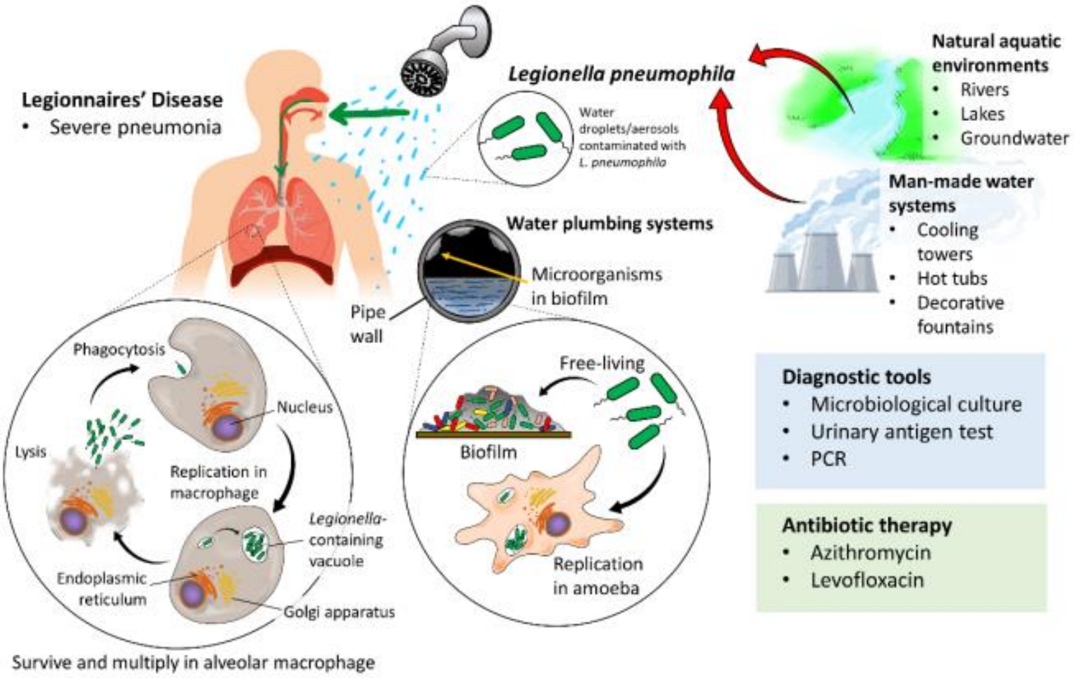


Figure 4 - Legionella pneumophila life cycle. From (Tan, Tee, Khan, Ming, & Letchumanan, 2021).

Pontiac's fever – The illness was first recognized in Pontiac, Michigan, USA, in 1968. It commonly manifests as a respiratory infection lasting two to five days, without advancing to pneumonia. Those affected may experience symptoms like fever, chills, malaise, muscle pain, headaches, coughing, or chest discomfort, typically after an incubation period of one to two days (Mraz & Weir, 2018).

Legionnaires' disease (LD) – Legionnaires' disease shares symptoms with Pontiac Fever but also involves more severe pneumonia. It can result in long-term effects such as ongoing fatigue, difficulty breathing, muscle or joint pain, and a productive cough (Mraz & Weir, 2018). Unlike the majority of individuals with community-acquired pneumonia, patients with Legionnaires' disease often exhibit gastrointestinal symptoms, as well as changes in mental status and neurological issues (NASEM, 2020).

The reasons why some patients develop Pontiac fever rather than pneumonia remain unclear, leaving the disease's pathogenesis uncertain. Several theories have been suggested, including the involvement of bacterial toxins, allergic responses, and interactions with amoebae that harbor *Legionella* (Edelstein, 2007). Because the symptoms of both diseases mimic those of other infections, diagnosing legionellosis may be delayed or missed, which could lead to serious complications for those with *Legionella* pneumonia. The groups most vulnerable to infection include smokers and individuals who work more than 40 hours per week. Additionally, the elderly, men, smokers, and those with hospital-acquired infections, kidney disease, or weakened immune systems are at a greater risk of fatal outcomes from pneumonia (Mraz & Weir, 2018). The weakening of both innate and adaptive immunity, known as immune senescence, significantly raises the risk among older people (NASEM, 2020).

2.5. Legislation related to *Legionella* in water systems in Portugal and Europe

The following table resumes some of the most recent national and European laws and guidelines related to *Legionella* control and prevention.

Table 2 – National and European Legislation related to Legionella.

Name Of the Law/Norm	
Lei Nº 52/2018 of 20 August	Establishes the legal framework for the prevention and control of *Legionella* in water systems, requiring monitoring, risk assessments, and maintenance protocols in high-risk environments such as hospitals, hotels, and industrial facilities.
Circular Normativa Nº 06/DT of 22/04/04	Provides technical guidelines for the prevention and control of *Legionella* in public buildings, focusing on surveillance, water treatment, and regular cleaning of water systems.
Orientação Nº 20/2017	Offers guidance for the investigation and management of *Legionella* outbreaks, defining steps for identifying sources, implementing control measures, and reporting cases to health authorities.
Norma Nº 024/2017	Defines specific actions for risk prevention of *Legionella* in healthcare environments, outlining maintenance

	procedures for water systems and medical equipment that uses water.
Decreto-Lei N.º 69/2023 of 21 August	Updates regulations on water system management for the prevention of *Legionella* contamination, with stricter penalties for non-compliance, enhanced monitoring requirements, and new obligations for building managers.
Directive (EU) 2020/2184 Of the European Parliament and Of the Council	Relates to the quality of water intended for human consumption, setting standards for microbiological, chemical, and indicator parameters in water, including *Legionella* monitoring, and outlines the responsibilities of Member States for ensuring compliance.

2.6. Climate change and *Legionella spp.*

Changes in climate, including increased rainfall and rising global temperatures, have been associated with a higher occurrence of diseases, either directly or through increased exposure to water sources connected with legionellosis (NASEM, 2020). As temperatures rise and certain regions face more precipitation, flooding, and elevated sea levels, concerns grow that *Legionella,* and other waterborne infections will become more widespread.

One of the main factors contributing to the increased risk of *Legionella* contamination is the rise in water temperatures. *Legionella* bacteria flourish in warm water, and as global temperatures rise, water systems become more prone to contamination. Longer warm seasons allow water systems to stay within *Legionella's* ideal growth range for extended periods, raising the risk of bacterial proliferation. Cooling towers, hot water systems, and plumbing infrastructure are especially at risk under these conditions (Walker, 2018).

Shifts in precipitation patterns, including heavier rainfall and increased humidity, have also been linked to higher rates of *Legionella* infection. Rainfall increases relative humidity, providing favorable conditions for the bacteria's survival and aiding its transmission via aerosols from sources like cooling towers, fountains, and showers. Research from the Netherlands revealed that the combination of warm weather and heavy rainfall greatly elevated the risk of community-acquired *Legionella* infections (Walker, 2018). Additionally, floods and heavy rains can disrupt water systems, leading to stagnation or contamination, both of which encourage bacterial growth.

An additional effect of climate change is the increased use of water recycling and rainwater harvesting, especially in regions affected by drought. While these systems offer essential water supplies, they also carry a risk of *Legionella* contamination if not adequately maintained.

Stored water can warm to temperatures that encourage bacterial growth, and activities such as pressure washing heighten the risk of aerosolization and inhalation exposure (Walker, 2018).

Given these environmental changes, public health strategies must adapt by enforcing stricter monitoring and management of water systems to prevent future outbreaks.

3. Methods

This review was based on the guideline of *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) methodology with the objetive of having a more transparent, complete and accurate report of this systematic review.

3.1. Eligibility Criteria

For this systematic review of predictive models of *Legionella*, the eligibility criteria were defined by the SPIDER (Sample; Phenomenon of Interest; Design; Evaluation; Research Type) methodology. The SPIDER methodology is particularly suited for designing the eligibility criteria in a systematic review focused on predictive models of *Legionella* due to its flexibility in accommodating a wide range of study designs and phenomena. By emphasizing key elements such as the Sample (the environments or populations where *Legionella* predictions are applied), the Phenomenon of Interest, and the Design (various study methodologies, including validation studies and comparative analyses), SPIDER allows for a comprehensive inclusion of both qualitative and quantitative research. The Evaluation component ensures that the review will critically assess the models' performance metrics, such as accuracy and precision. At the same time, the Research Type broadens the scope to incorporate diverse types of studies that contribute to understanding how these models function in different contexts. This makes SPIDER an optimal framework for capturing the multifaceted nature of predictive modeling in public health settings like *Legionella* detection. Therefore, the inclusion and exclusion criteria were defined and are detailed in Table 3. Publications in Portuguese, English and Spanish were also included, from the last ten years and with complete and open access. The exclusion criteria included publications in languages other than English, Portuguese, Spanish, or Italian, those more than 10 years old, publications with restricted access, as well as systematic reviews, meta-analyses, and organizational reports.

Table 3 – Eligibility criteria for study inclusion.

	Inclusion criteria	**Exclusion criteria**
Sample	Hospitals, hotels, bathrooms and every at-risk area for *Legionella* infection.	Areas not identified as at risk for *Legionella* infection.
Phenomenon of Interest	Detection and prediction of *Legionella spp.*	The objective of study different from the stated.

Design	Studies that develop and test a predictive model for *Legionella* contamination.	Studies that focus solely on identifying variables or parameters without testing or validating a predictive model.
Evaluation	The study must test the performance of the predictive model, providing results that evaluate its effectiveness.	Studies where the model is proposed but not tested or validated in real-world or simulated scenarios.
Research Type	Observational studies.	Systematic reviews, meta-analyses and reports.
Other characteristics	Publications in English, Portuguese, Spanish, and Italian; from the last 10 years (2014–2024); open access or accessible through institutional subscriptions.	Publications in other languages not included in the review or older than 10 years; restricted access without institutional or open access availability.

3.2. Information Sources, Search Strategy and Selection Process

Information sources – the research of the publications in this review was conducted in three databases: PubMed; ScienceDirect and Web of Science. The publications included were dated between the 1st of January of 2014 and 31st of August of 2024.

The keywords used in the research of the articles were: "*legionella*", "*legionella pneumophila*", "predictive", "prediction", "model" and "aerosol". The boolean operators "AND" and "OR" were used to combine the different keywords and to achieve the best results possible. Filters were also applied in the databases during the research of the articles, those were: year of publication, language (Portuguese, Spanish and English), free full text and the exclusion of other systematic reviews, meta-analysis and reports.

Research question – Which model is more reliable for the prediction and prevention of *Legionella* infections?

Research and study selection – After searching for articles in the various databases, these were imported into Mendeley Reference Manager 2.79.0—a free reference management software—thus eliminating duplicate articles. After removing duplicate articles, two reviewers (MS and HC) conducted, independently, the article selection in two phases. In the first phase, the titles and abstracts were analysed, and studies were excluded based on the previously established eligibility criteria. In the second phase, the full texts of the studies were read, analyzed, and selected for subsequent data extraction.

3.3. Data Collection Process and Data Items

The data from the studies included in this systematic review were extracted after a full reading. The data were meticulously organized in Table 5, with the information categorized as follows: author(s), year of publication, country, model, study objective, sample, evaluation, phenomenon of interest, and study limitations. Due to the diversity of the studies, a statistical analysis of the results was not conducted. Instead, a qualitative analysis of the studies was performed.

3.4. Evaluation of Methodology Quality and Study Risk of Bias Assessment

The studies' methodological quality and risk of bias were evaluated using the Checklist for Systematic Reviews and Research Syntheses (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003). The Joanna Briggs Institute (JBI) is an internationally respected research organization based at the University of Adelaide, Australia, with a focus on evidence-based healthcare. JBI offers extensive guidelines and methodologies for conducting systematic reviews and research syntheses across multiple healthcare and public health disciplines. Its tools are well-known for their effectiveness in assessing the quality and rigor of primary research, ensuring that systematic reviews produce reliable and clinically meaningful results. JBI is widely used to evaluate the methodological quality of studies, minimize bias, and enhance the transparency and reproducibility of systematic review (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003). In the context of this systematic review on predictive modeling for *Legionella*, the JBI tools are integral to maintaining a high standard of evidence synthesis. JBI's critical appraisal checklists allow researchers to assess key factors such as the study's internal validity, the adequacy of sample sizes, the appropriateness of statistical analysis, and how well the study addresses potential sources of bias. By doing so, the systematic review can more confidently include studies that meet high-quality standards and exclude those that may compromise the reliability of the results (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003).

JBI's tools are particularly valuable for systematic reviews in public health domains, such as this one on *Legionella*, because they offer a structured approach for reviewing studies across a wide range of research designs. The complexity of *Legionella* contamination and its environmental and health impacts require models that are based on solid empirical evidence. JBI helps to ensure that the predictive models reviewed in this study are assessed with rigor, allowing the systematic review to present findings that are both scientifically robust and practically relevant. Additionally, the use of JBI tools enhances the overall transparency of the review process, making it easier for future researchers to replicate or build upon the finding (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003).

Thus, the JBI methodology ensures that the systematic review on *Legionella* predictive modeling incorporates only the most reliable studies, improving the accuracy and utility of the predictive models being reviewed. This, in turn, facilitates better public health interventions and more effective management of *Legionella* contamination risks. The robust framework provided by JBI is key to achieving the objectives of this review, particularly in producing high-quality, actionable insights in the field of environmental health and disease prevention.

The Checklist for Systematic Reviews and Research Syntheses is provided in Annex A, and the results from the analysis of the articles in this review are presented in Annex B.

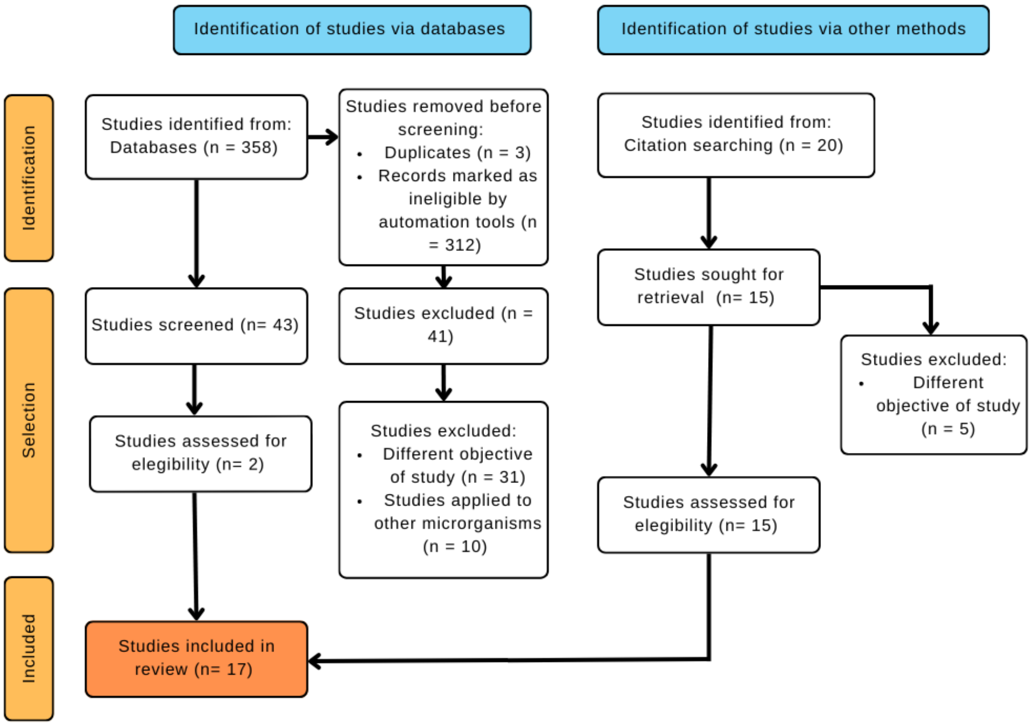


Figure 5 – Study selection flowchart, adapted from PRISMA methodology (Page, et al., 2021).

4. Results

After the research strategy was applied, a total of 358 articles were retrieved from the three databases for this review. Of these, 315 were excluded: 312 were identified as ineligible by automation tools, and 3 were duplicates. This left 43 articles for further assessment, from which 41 were excluded—31 due to differing study objectives and 10 because they focused on organisms other than *Legionella*. The remaining 2 studies were fully read and analyzed for eligibility. Additionally, a citation search was conducted, resulting in the inclusion of 15 more studies that were relevant to the focus of this review. In conclusion, a total of 17 articles were included in this study for qualitative analysis.

4.1. Included Studies

Starting from the general objective of this systematic review, the evaluation and limitations of the models studied are included in the data presented in Table 5 below, structured to provide a more comprehensive answer to the research question. Consequently, the results are organized as follows: author/year of publication, country, model, study objective, sample, study design, evaluation, phenomenon of interest, and study limitations.

4.2. General Characteristics of Studies

The studies analysed in this systematic review primarily focus on developing and refining predictive models to assess the risk of *Legionella* contamination in various water systems.

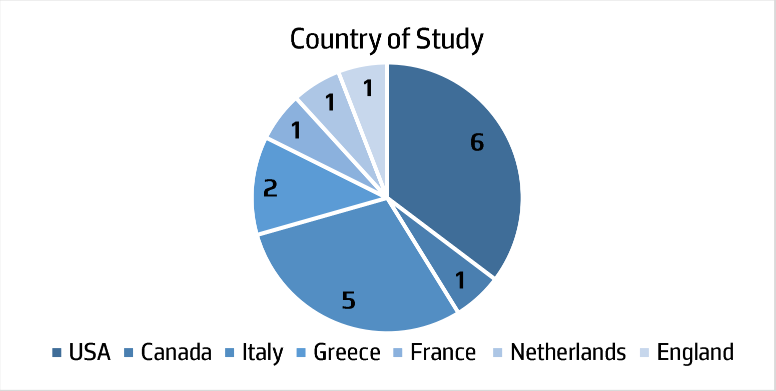


Figure 6 – Countries of origin for the included studies.

These studies are conducted across different countries, focusing on both natural and built environments, such as hospitals, hotels, cooling towers, agricultural irrigation systems, and residential buildings.

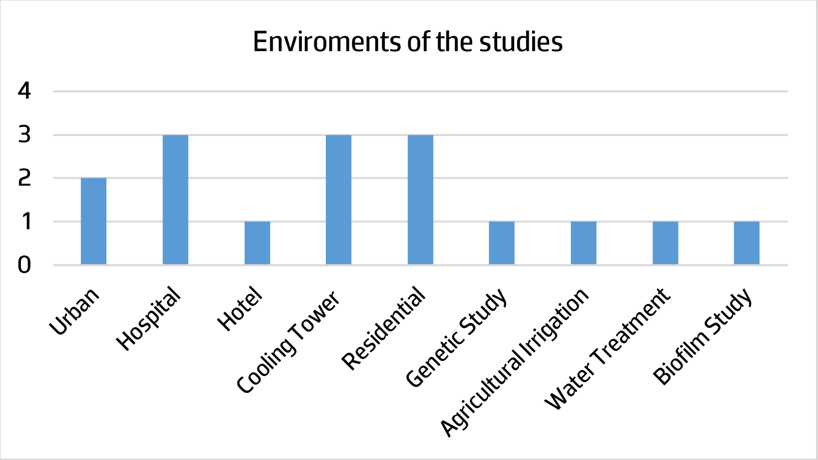


Figure 7 – Bar chart showing the different types of environments where the studies were conducted.

Across the studies, several common variables are consistently identified as significant contributors to *Legionella* growth, including water temperature, chlorine levels, biofilm presence, and environmental conditions such as humidity and particulate matter (Chochlakis et al. 2023;Dye et. al 2021; De Giglio et al. 2024). Some studies focus on infrastructure-specific variables, such as water system design or filter maintenance, which have also been found to significantly impact *Legionella* contamination levels (De Giglio et al. 2021; H. Weir et al. 2019).

Table 4 - Important factors for Legionella proliferation and maximum contamination reported in each study.

Study	**Important variables/risk factors for *Legionella* growth**	**Maximum Contamination Levels reported**
Hancock et al. (2014)	Human movement patterns, presence of cooling towers, urban density.	No maximum contamination levels reported.
Brunello et al. (2022)	Temperature, humidity, geographic factors, water system design.	No maximum contamination levels reported.
Chochlakis et al. (2023)	Water temperature (higher temperatures favored growth), low chlorine levels, co-presence of *Pseudomonas aeruginosa*.	Maximum contamination of 10,000 CFU/L in some samples.
De Giglio et al. (2021)	Building height (number of floors), water system design (complexity), water stagnation.	No maximum contamination levels reported.
De Giglio et al. (2024)	Temperature fluctuations, chlorine level fluctuations, filter maintenance practices.	Maximum contamination of 3,000 CFU/L detected in hospital systems.
Dye et al. (2021)	Chlorine levels, water temperature, pH levels, system stagnation.	No maximum contamination levels reported.
Girolamini et al. (2023)	Meteorological conditions, particulate matter (PM_{10} levels), water temperature.	No maximum contamination levels reported.
Hamilton et al. (2019)	Exposure from residential fixtures (e.g., showers).	Simulated contamination levels ranged from 1,000 to 5,000 CFU/L.
Massiot et al. (2023)	Water temperature, irrigation system design.	Maximum contamination reported as 15,000 CFU/L in wastewater.
Papagianeli et al. (2021)	Water temperature.	Maximum contamination of 12,000 CFU/L after temperature control.
Ortiz et al. (2024)	Biofilm formation, water temperature, nutrient availability, chlorine concentration.	Biofilm contamination levels reached up to 20,000 CFU/L.
Spica et al. (2024)	Building infrastructure, chlorine levels.	Maximum contamination reported at 2,500 CFU/L in residential areas.
Xu et al. (2024)	Amoeba-host interactions, water pH, water temperature, chlorine and nutrient levels.	No maximum contamination levels reported.

Additionally, certain studies report on maximum contamination levels, typically expressed in colony-forming units (CFU) per liter, providing valuable insight into the severity of contamination under different conditions (Chochlakis et al. 2023; De Giglio et al. 2021; Girolamini et al. 2023).

Most of the studies utilize a range of advanced modeling techniques, including machine learning, regression models, and mechanistic models, to gain deeper insights into the factors driving *Legionella* proliferation (Dye et. al 2021; De Giglio et al. 2024; Chochlakis et al. 2023). These approaches allow researchers to analyse complex datasets and identify critical variables influencing contamination, enhancing the predictive capabilities of these models in various water systems (De Giglio et al. 2021; H. Weir et al. 2019).

The variety of study designs highlights the complexity of managing *Legionella* contamination, as multiple factors — environmental, infrastructural, and biological — interact to create contamination risks.

4.3. Description and Analysis of Results

The studies included in this systematic review take diverse approaches to predicting and mitigating *Legionella* contamination in various water systems. Each study focuses on different environmental settings, from hospitals and cooling towers to residential areas and agricultural irrigation systems, employing a variety of methodologies such as machine learning, risk assessments, and regression models. Below is a more detailed description of each study, focusing on its objectives, methods, findings, and relevance.

The predictive model developed by **Hancock et al. (2014)** is a mathematical model designed to estimate the dynamics of non-transmissible infections like *Legionella pneumophila*, focusing on outbreaks caused by localized environmental sources. The model is built upon detailed human movement patterns across Great Britain, sourced from the Great Britain Human Movement (GBHM) database. This database includes data from commercial sources on travel behaviour for work, shopping, and education, incorporating socio-demographic and geographic variables to predict human movement.

The model calculates the probability of infection by relating human movement data to a potential infection source, which is represented by hexagonal spatial units. The probability of becoming infected by a source is calculated using the formula:

$$p(i|S,\lambda_s,\beta_s)=1-\exp\left(-\lambda_s\tau\sum_{d\in D}(P^d_{i,v}(S)+\beta_s P^d_{i,h}(S))\right)$$

Where:

- λ_s represents the infection rate in the source hexagon.
- β_s is a constant representing the infection risk while at home.
- $P^d_{i,v}(S)$ and $P^d_{i,h}(S)$ are probabilities that individual i is present in location S (outside of home) or at home during day part d, respectively.
- t is the time.

This infection probability is used to determine the likelihood of infection in a population over a set period, and the model estimates which spatial units (hexagons) are likely to contain the infection source.

The most important variables in this study were:

- **Human Movement Patterns**: Based on work, shopping, and educational activities, stratified by socio-demographics.
- **Distance from Source**: Travel distance plays a crucial role in determining exposure risk.
- **Environmental Conditions**: Urban geography and attractiveness of locations (e.g., shopping centres) are integrated into the model to improve prediction accuracy.
- **Population Data**: Census and commercial travel data, age, gender, and employment status are considered in determining movement probabilities.

This model successfully predicts the location of the infection source and identifies high-risk individuals based on their proximity to the source and movement patterns, providing a robust framework for public health intervention strategies.

Dye et al. (2021) uses a neural network model developed with the TensorFlow machine learning platform to predict the probability of *Legionella* contamination in water samples from premise plumbing systems. The model was trained using physicochemical water quality parameters such as total chlorine, free chlorine, temperature, and pH. It aimed to determine whether these parameters correlated with a detectable level of *L. pneumophila* in water samples above the threshold of 10 MPN (Most Probable Number) per 100 mL. The neural network used in this study consists of an input layer, two hidden layers (each with 32 nodes using the ReLU activation function), and a final dense layer (with one node using a sigmoid activation function). The model was trained to classify whether a sample contained *Legionella* at detectable levels. The network was trained using the binary cross-entropy loss function and optimized with the Adam optimizer at a learning rate of 0.001.

The most important variables in this study were: total chlorine, free chlorine, temperature and pH levels. These variables were systematically trained and evaluated, and the neural network was able to identify water samples with detectable *Legionella* contamination with a high level of accuracy. The model achieved an F1 score of 0.90, demonstrating its capability in predicting the presence of *Legionella* under different water quality conditions.

The study by **Brunello et al. (2022)** employs machine learning techniques to analyse the spatiotemporal spread of *Legionella* bacteria using geocoded data. The primary model used in this analysis is a tree-based ensemble classification method, specifically XGBoostClassifier, which builds decision trees for classification. This model was used to predict *Legionella* contamination levels and serogroups based on several environmental factors, including climatic and geographic data. The XGBoost model is a form of gradient boosting decision trees, which iteratively builds models to predict residual errors from previous iterations. By minimizing loss functions through a gradient descent algorithm, the model combines the strengths of decision trees into an ensemble, which produces more accurate and robust predictions. The most important variables in this study were:

- **Water Temperature**: *Legionella* growth was favoured by temperatures between 25°C and 42°C.
- **Humidity**: Humidity levels, particularly below 58% and above 91%, were crucial in predicting the presence of *Legionella*, especially the more virulent *L. pneumophila* serogroup 1.
- **Latitude and Longitude:** These spatial features were also significant, helping determine the risk zones where *Legionella* contamination was most likely to occur.
- **Rainfall and Temporal Features:** Rainfall and day of the year were integrated to account for the seasonal variations in *Legionella* contamination.

The Shapley values method was applied to interpret the contribution of each variable in the model, explaining how various environmental factors contributed to the model's predictions.

Papagianeli et al. (2021) uses an inactivation kinetic model to predict the thermal inactivation of *L. pneumophila* in water systems. The primary focus is on the effect of water temperature in reducing *L. pneumophila* populations under isothermal and dynamic temperature conditions. The study developed a log-linear model to describe the inactivation kinetics of *L. pneumophila* in water, focusing on temperature-dependent inactivation processes. The primary model used is the Bigelow model, which calculates the D-values (decimal reduction time), representing the time needed for a 1-log reduction of the microbial population at a constant temperature.

The model is expressed as:

$$\log_{10} N = \log_{10} N_0 - \frac{t}{D_T}$$

Where:

- N_0 is the initial number of microorganisms.
- N is the number of microorganisms after time t.
- D_T is the time required for a 1-log reduction at temperature T.

A secondary model, known as the z-value model, was used to describe the effect of temperature on D-values, with the formula:

$$\log_{10} D_T = \log_{10} D_{ref} + \frac{T_{ref} - T}{z}$$

- D_{ref} is the D-value at the reference temperature T_{ref}.
- Z is the temperature increase required for a 10-fold reduction in D-value.

The most important variables in this study were:

- **Water temperature (°C):** The study tested isothermal conditions ranging from 51°C to 61°C and dynamic profiles mimicking real-world water thermal disinfection conditions.
- **Inactivation time (minutes):** The model calculates how long the bacteria remain viable at different temperatures.

The model was validated with dynamic temperature conditions to predict *L. pneumophila* inactivation across various water systems, indicating it can translate temperature changes into bacterial reduction in real-time.

The study by **Girolamini et al. (2023)** uses a quasi-Gaussian dispersion model known as the ADMS-Urban model. This predictive model is designed to simulate the spread of bioaerosols containing *Legionella* from cooling towers (CTs) in urban environments. The primary focus is on predicting the spatial dispersion of *Legionella*-contaminated aerosols, using PM_{10} particles as a proxy for potential *Legionella* transmission.

The ADMS-Urban model is a widely used dispersion model that simulates the behaviour of pollutants emitted from various sources, including cooling towers. The model calculates how aerosols, such as those potentially containing *Legionella*, are transported and dispersed through the atmosphere based on local meteorological conditions (e.g. wind speed, temperature, humidity).

In this study, the model used PM_{10} (particulate matter less than 10 µm in diameter) as a proxy for bioaerosol transmission. PM_{10} was chosen because *Legionella* bacteria are often transmitted through aerosols of this size. The model accounted for various factors such as: tower height, volumetric flux of release, air temperature, wind speed and direction and humidity levels. The model calculates the risk of *Legionella* spreading in different weather conditions, focusing on the distance that contaminated aerosols could travel from the cooling tower. The key variables considered in this model include:

- **Air temperature (°C):** Higher temperatures can increase the viability of *Legionella* bacteria in the aerosols.
- **Wind speed and direction**: These factors influence how far and in which direction the aerosols are transported.
- **Humidity levels:** High humidity can enhance the survival of the bacteria in the aerosols.
- **Precipitation (mm/day):** Precipitation events prior to sampling were considered significant in *Legionella* dispersion.

The study concluded that the risk of Legionella spread was highest in winter and summer, with the potential for contaminated aerosols to travel up to 1,5 km from the cooling tower under certain weather conditions. This model provides a predictive tool for assessing the risk of *Legionella* spread in urban environments.

The study by **Wong et al. (2024)** uses a two-stage deep learning model to automatically detect cooling towers from aerial imagery. The first stage employs the YOLOv5 (You Only Look Once version 5) object detection model to locate cooling towers, and the second stage uses EfficientNet-b5, an image classification model, to further verify the detected cooling towers. This model was designed to enhance the efficiency of Legionnaires' disease outbreak investigations by identifying cooling towers that might serve as sources of *Legionella* contamination. The model operates in two stages:

- Stage 1: YOLOv5 – This is an object detection model that scans aerial imagery to detect objects that could be cooling towers. YOLOv5 assigns a probability score to each detected object, with thresholds to classify them as high, intermediate, or low probability.

- Stage 2: EfficientNet-b5 – For objects identified with intermediate probability in stage 1, EfficientNet-b5, a deep learning-based image classification model, is used to distinguish between actual cooling towers and false positives.

The YOLOv5 model was trained using manually annotated images from New York and Philadelphia and expanded with synthetic data to improve generalization to other cities. EfficientNet-b5 was used to crop and evaluate detected objects, improving the positive predictive value (PPV) of the model. The most important variables and inputs in this model include:

- **Aerial images:** The model uses satellite view images from Google Maps, and the system processes 1280x1280 pixels.
- **Object features:** The model identifies geometric features typically associated with cooling towers, such as fan shapes and building positions.
- **Synthetic data**: The training data included synthetic images generated to simulate a variety of urban landscapes.

The model achieved a sensitivity of 95.1% in New York and Philadelphia and was more than 600 times faster than manual searches.

De Giglio et al. (2021), another study from Italy, aimed to develop a scoring tool to predict *Legionella* contamination in recreational and touristic facilities. This study used data collected over 18 years from 47 facilities and a Poisson regression model to assess *Legionella* contamination risk in tourist facilities. This model focuses on three main outcomes: water samples positive for *Legionella*, water samples exceeding 1000 CFU/L of *Legionella*, and clinical cases of Legionnaires' disease. The Poisson regression model formula is given by:

$$e^{(\alpha+\beta 1X1+\beta 2X2+\ldots+\beta kXk)},$$

Where:

- α - intercept of the model.
- β - regression coefficients for each risk factor.
- X - scores assigned to each parameter (risk factor).

This formula calculates a risk score based on several structural and environmental parameters (e.g., facility type, water system features, air conditioning systems) that influence *Legionella* proliferation. The model was validated using a large dataset collected over 18 years from clinical and environmental surveillance. The resulting risk scores were then used to predict the likelihood of *Legionella* contamination or disease occurrence.

The model included 47 different parameters related to water systems, air conditioning, and building features. Key variables include water temperature, biofilm formation, facility size, number of showers, and the presence of wellness services like saunas and whirlpools.

De Giglio et al. (2024) focuses on predicting Legionella contamination in a hospital water network using a combination of machine learning (ML) models and Poisson regression analysis. The primary goal was to evaluate which environmental, structural, and maintenance-related factors could be used to predict the presence of *Legionella* in water samples.

The machine learning model used in this study was a supervised neural network designed for binary classification (positive or negative *Legionella* detection). It consisted of:

- Input Layer: 57 input variables representing structural, environmental, and maintenance parameters.
- Hidden Layer: One hidden layer with 30 neurons, using the Rectified Linear Unit (ReLU) activation function.
- Output Layer: Two neurons, classifying water samples as either positive or negative for *Legionella*.

The machine learning model was trained using 70% of the dataset, while 30% was used for testing. The parameters influencing the detection of *Legionella* included water temperature, type of water network (hot/cold), filter valve replacement history, and atmospheric conditions.

The Poisson regression model was employed to predict the *Legionella* load (cfu/L) in positive samples. The model assessed the relationship between various independent variables (structural and environmental) and the dependent variable (*Legionella* contamination levels).

The Poisson regression formula is given by:

$$e^{(\alpha+\beta 1X1+\ \beta 2X2+\ldots+\ \beta kXk)},$$

Where:

- α – intercept of the model.
- β – regression coefficients for each risk factor.
- X – scores assigned to each parameter (risk factor).

The most important variables in this study were:

- **Water temperature:** Higher water temperatures were associated with increased *Legionella* contamination.
- **Type of water network (hot/cold):** Cold water networks were found to have a higher risk of contamination.
- **Atmospheric conditions:** Warmer daily temperatures correlated with higher contamination rates.
- **Maintenance practices**: The frequency of filter valve replacements was a significant predictor of contamination, with the risk increasing as the time since the last replacement increased.

The study concluded that the machine learning model outperformed the Poisson regression model in terms of prediction accuracy, achieving an accuracy of 93.4%. However, the Poisson regression model provided important insights into the factors influencing *Legionella* load, complementing the machine learning model's predictions. Combining both approaches allowed for more robust predictions, particularly in assessing contamination risk in hospital water systems.

Chochlakis et al. (2023) in Greece used a generalized linear model (GLM) combined with a binomial logistic regression model to assess the likelihood of *Legionella* presence in hotel water systems. The model was designed to estimate the probability of detecting *Legionella* based on various environmental, microbiological, and physicochemical factors, including water temperature, chlorine concentration, and the presence of other pathogens such as *Escherichia coli* and *Pseudomonas*.

In the context of the model, the formula calculates the log-odds of *Legionella* presence in relation to the independent variables. For example, the relationship between water temperature and *Legionella* detection is modeled as either increasing or decreasing the probability of finding *Legionella* based on temperature values. In cold water systems, the model identified a higher likelihood of *Legionella* presence, while in hot water systems, the likelihood decreased. Additionally, the formula incorporates other variables like chlorine concentration, which has a mitigating effect on bacterial contamination.

The logistic regression approach was particularly useful for assessing binary outcomes—contamination or no contamination — by estimating how changes in factors like temperature and chlorine concentration affect the probability of *Legionella* growth. The study highlighted temperature and microbial interactions as significant predictors, providing insight into how hotel water systems can be managed to reduce *Legionella* risks.

H. Weir et al. (2019) employs a Quantitative Microbial Risk Assessment (QMRA) model to estimate the risk of legionellosis within diverse populations. QMRA is a computational science approach used to forecast microbial infection risks, often applied in environmental health and public health contexts. In this case, the model focuses on *L. pneumophila*, considering factors like biofilm presence, water system contamination, and population diversity in susceptibility to illness. The model integrates population-specific variables, such as age, sex, and race, to enhance the accuracy of risk predictions for distinct demographic groups. It uses dose-response modeling to estimate infection risks, accounting for various exposure pathways, such as inhalation of aerosolized water droplets in showers. The model incorporates stochastic simulations (Monte Carlo method) to account for uncertainties in water concentrations and exposure durations, refining the accuracy of risk estimations. Variables that determine the model include water temperature, biofilm presence, water flow rate, concentration of *Legionella* in water systems, aerosol generation, and demographic factors such as age and sex. The model improves upon traditional QMRA methods by refining the representation of population-specific risks, making it applicable to more diverse and vulnerable groups (such as the elderly and immunocompromised).

The study by **Hamilton et al. (2019)** uses a Quantitative Microbial Risk Assessment (QMRA) model to assess the risk of *L. pneumophila* exposure in indoor residential water systems, specifically focusing on aerosol-producing fixtures like showers, toilets, and faucets. The QMRA model estimates the probability of infection or clinical severity infection (CSI) based on various exposure scenarios and environmental variables. The basis of this model involves calculating the exposure to *Legionella* aerosols by applying dose-response functions to determine the risk of infection. Two primary dose-response models are used: one for sub-clinical infection and one for clinical severity infection. The QMRA framework also factors in the variability of exposure based on different water fixtures, considering aerosol size, concentration, and partitioning from water to air.

Key variables include:

- **Water fixture aerosol concentration:** The concentration of *Legionella* in water and the fraction of *Legionella* that becomes aerosolized.

- **Aerosol size distribution:** Different fixtures produce aerosols of varying sizes, and aerosols within the 1–10 µm size range are most likely to be inhaled and deposit in the alveoli, posing a risk of infection.
- **Breathing rate and exposure time:** The volume of air inhaled and the time spent near water fixtures contribute to the total exposure.
- **Dose-response parameters:** These define the relationship between the exposure dose and the probability of infection or clinical severity, with parameters for both mild and severe infections.

The QMRA approach used by Hamilton et al. provides an estimate of critical *Legionella* concentrations required to stay below specific health risk targets, like infection or DALY-based (disability-adjusted life year) targets, which are used to inform public health guidelines for water system management.

In **Massiot et al. (2023)** the General Bayesian Network (GBN) model is employed as a predictive tool to monitor the risk of *Legionella* infection during agricultural irrigation using reclaimed wastewater. The GBN integrates quantitative microbial risk assessment (QMRA) methodologies, focusing on exposure to *L. pneumophila* through aerosolized water droplets. The GBN model is designed to work with both discrete and continuous variables, making it highly adaptable to variations in data availability and environmental conditions. The model uses a directed acyclic graph (DAG) to represent the exposure pathway from the wastewater treatment plant (WWTP) to the population. The exposure pathways are influenced by variables such as the concentration of *Legionella* in the wastewater, water treatment efficiency, weather conditions, and the proximity of individuals to the irrigation source. The model updates predictions based on new data inputs, enabling dynamic risk assessment as additional field measurements become available. Key variables in this model include *Legionella* concentrations in water, the decay rates of the bacteria during different treatment processes (ultra-filtration and UV treatments), and meteorological factors such as wind speed and direction, which affect the dispersion of aerosolized *Legionella*.

Additionally, the model accounts for different population exposure scenarios (passersby, residents, and farmers) and calculates the risk of infection over time. The GBN integrates multiple components: water contamination (measured at different points in the treatment process), atmospheric dispersion (a modified Gaussian plume model is used to predict bacterial concentrations in the air), and exposure (the inhaled dose based on individual activities). By linking these elements, the GBN model can assess both short-term and long-term risks of infection, providing a comprehensive framework for managing *Legionella* risks in agricultural settings. The GBN model operates within the QMRA framework. The model allows for continuous updating of risk estimates based on new data, utilizing Bayesian inference to adjust predictions as more information becomes available. This flexibility is a major strength of the GBN approach, particularly in scenarios with uncertain or limited data. The formulas governing the model include:

- Logarithmic decay formulas for bacterial concentration during water treatment.
- Gaussian plume dispersion model for airborne *Legionella* transmission.
- Dose-response model to calculate infection probabilities based on inhaled doses.

These elements are combined to predict the likelihood of *Legionella* infection in different exposure scenarios, considering variations in water treatment efficacy and environmental conditions. Important variables in this study:

- Water contamination levels at different stages of wastewater treatment.
- Decay rates for *Legionella* during filtration and UV treatment.
- Meteorological factors such as wind speed, direction, and temperature.
- Proximity of individuals to the irrigation source.
- Exposure scenarios, including the duration and intensity of exposure for various population groups.

The study by **Van der Ploeg (2016)** compares different predictive models for classifying *L. pneumophila* strains as either clinical or environmental. The focus was on four modeling techniques: Classification and Regression Trees (CART), Random Forest (RF), Support Vector Machines (SVM), and Least Absolute Shrinkage and Selection Operator (LASSO). Each model aims to identify the most relevant genomic features that can differentiate between clinical and environmental strains of *Legionella*, using a dataset of 222 *Legionella* strains with 448 continuous markers.

- **Classification and Regression Trees (CART)**

CART is a decision tree-based model that uses recursive partitioning to classify data. The model identifies the best splits by minimizing an impurity index, splitting the data into segments until a stopping criterion is reached. CART is simple to interpret but can lead to overfitting, especially with small datasets.

- **Random Forest (RF)**

Random Forest is an ensemble model that builds multiple decision trees from random subsets of the dataset. It outputs the most frequent classification among all trees, making it more robust against overfitting compared to single–tree methods like CART. RF was found to have the highest predictive performance (AUC = 0.975) in this study.

- **Support Vector Machines (SVM)**

SVM constructs hyperplanes in a high-dimensional space to separate different classes. It is particularly effective in datasets with large numbers of predictor variables. The model maximizes the margin between the separating hyperplanes, ensuring better generalization. SVM showed a moderate performance with an AUC of 0.859 in this study.

- **LASSO (Least Absolute Shrinkage and Selection Operator)**

LASSO is a linear regression model that includes a penalty term, shrinking some coefficients to zero. This process enables feature selection, keeping only the most relevant variables in the model. LASSO was highly effective in this study, with a validated AUC of 0.925.

Each of these models used the same input data but processed the feature selection and classification differently. RF and LASSO emerged as the best models for identifying significant genomic markers. The performance of the models was measured using the Area Under the Receiver Operating Characteristic (AUC) curve. Key Variables in this study:

- The genomic markers were selected based on their frequency and relevance in differentiating between clinical and environmental strains. The most common genomic feature across all models was LePn.007B8, which was consistently identified as a significant marker for distinguishing clinical strains.

In summary, Random Forest and LASSO provided the highest performance in predicting *Legionella* strain classification, demonstrating the effectiveness of these techniques in genomic feature selection.

Wilson et al. (2022) focuses on the Quantitative Microbial Risk Assessment (QMRA) model. The QMRA model combines environmental monitoring data with human exposure scenarios to predict the risk of Legionnaires' disease. The model takes into account key variables such as the concentration of *L. pneumophila* in the water, aerosol generation during showering, and the deposition of these aerosols in the human respiratory system. The model compares two specific exposure models — Model 1 and Model 2 – both of which estimate infection risk from *Legionella* exposure during showering. The models consider several factors, including:

- Concentration of *Legionella* in the water.
- Aerosolization rates, which convert waterborne *Legionella* into airborne droplets.
- Inhalation rates and the amount of air breathed during a showering event.
- Deposition efficiency, which accounts for how much of the *Legionella* reaches the alveoli in the lungs.

The QMRA model is governed by dose-response relationships, which estimate the probability of infection given a certain dose of *Legionella* inhaled. Model 1 and Model 2 use different methods for partitioning *Legionella* in the air and its deposition in the lungs, but both eventually feed into a similar exponential dose-response curve, which helps predict infection probability.

The model calculates infection risk as a function of aerosol concentration, exposure time, and the individual's age, sex, and respiratory health. Model 2, which has more refined aerosol size distributions and conservative estimates, is highlighted as more reliable for app development in managing water safety in large public buildings. The principal variables used in the model include:

- Legionella concentration in water (CFU/mL).
- Shower duration.
- Aerosol size and deposition rate.
- Inhalation rate.

The model's ability to predict risks is improved by integrating demographic data, which allows for a population-specific risk analysis.

The study by **Ortiz et al. (2024)** uses a mechanistic model based on EPANET-MSX, an extension of the EPANET hydraulic and water quality simulation tool. This model is applied to predict *Legionella* concentrations within building water systems (BWS). EPANET-MSX allows the simulation of various physicochemical reactions and microbial growth within complex water distribution systems. In this study, the focus is on simulating the growth and transport of *Legionella spp.* within the water system, particularly the influence of temperature, chlorine levels, biofilm interactions, and nutrient availability on bacterial proliferation. The model considers four main factors: nutrient availability, temperature, chlorine concentration, and biofilm detachment. The model's framework sequentially integrates these factors to estimate *Legionella* growth and decay within the water system.

The study uses a first-order growth equation to model the bacterial growth in bulk water, taking into account both growth and mortality rates:

$$\frac{dC_{LEG,b}}{dt} = (\mu - k_d) \cdot C_{LEG,b}$$

Where:

- C_{leg} is the concentration of *Legionella* in the bulk water.
- μ is the growth rate of *Legionella*.
- K_d is the mortality rate of *Legionella*.

The model also includes the effect of temperature, using a Gaussian distribution to simulate how the bacteria respond to different temperature ranges. Chlorine concentration is integrated into the mortality rate (K_d), based on chlorine's disinfectant capabilities. Finally, biofilm interactions are modeled to simulate *Legionella* detachment from biofilms, which can significantly increase bacterial concentration in stagnant water. Key variables include:

- **Water temperature:** The model accounts for optimal *Legionella* growth between 25°C and 45°C, with mortality occurring at higher temperatures.
- **Chlorine levels**: Chlorine reduces *Legionella* concentrations in both the water and biofilm, influencing bacterial decay.
- **Biofilm detachment**: Biofilms can harbour *Legionella*, and detachment can release large amounts of bacteria into the water.
- **Nutrient availability:** The concentration of biodegradable dissolved organic carbon (BDOC) affects bacterial growth, with nutrient limitations slowing down *Legionella* proliferation.

This model provides insights into how complex factors within building water systems interact to influence *Legionella* concentrations, emphasizing the importance of biofilm dynamics and chlorine disinfection for controlling bacterial growth.

Spica et al. (2024) presents a mathematical model developed to assess the effectiveness and efficiency of different *Legionella* monitoring and control strategies in Italy. The purpose of the model is to inform public health policies by comparing the economic and human costs associated with two different approaches: targeting *L. pneumophila* or *Legionella spp*. more broadly.

The model is grounded in cost-benefit analysis and Quantitative Microbial Risk Assessment (QMRA), which is used to estimate the risks associated with exposure to *Legionella* in various building types. By combining epidemiological, economic, and environmental data, the model simulates different real-life scenarios, such as water contamination levels and varying building environments. The data input parameters include factors like building water temperature, *Legionella* prevalence, and pathogen virulence. Key variables that govern the model include:

- Water system contamination levels, measured in colony-forming units (CFU) per liter.
- Building water temperature, which affects *Legionella* growth.
- Species differentiation, focusing on the prevalence of *L. pneumophila* vs. other *Legionella* species.
- Human health outcomes, using years of life lost (YLL) and years lived with disability (YLD) to assess the burden of disease.
- Economic costs, including costs of building remediation, legal expenses, and productivity losses from temporary building closures.

The model predicts that controlling for *L. pneumophila* specifically is both cost-effective and provides significant health benefits, preventing Legionnaires' disease cases more efficiently than a broader *Legionella* control approach.

The study by **Xu et al. (2024)** develops a machine learning (ML) model aimed at low-cost real-time monitoring of *Legionella* concentrations in premise plumbing systems. The model is grounded in statistical learning theory and employs multiple methods, including LASSO (Least Absolute Shrinkage and Selection Operator) and multivariate linear regression (MLR). The model is built on the premise that *Legionella* contamination can be predicted through water quality indicators and the presence of host cells, such as free-living amoeba (FLA). The authors first use LASSO and MLR to assess the impact of physiochemical parameters (e.g. temperature, pH, free chlorine) on FLA concentrations. They then apply logistic regression to quantify the likelihood of *Legionella* contamination based on the concentrations of FLA and environmental indicators. The key variables in this study include temperature, pH, total chlorine (TCL), free chlorine (FCL), and total dissolved solids (TDS). The results demonstrate that the concentration of *Legionella* correlates strongly with the presence of amoebae like *Naegleria fowleri* and *Acanthamoeba spp.,* both of which serve as host cells. This model provides a robust method for estimating *Legionella* concentration ranges using physiochemical parameters and host cell concentrations, which makes it valuable for real-time risk monitoring in building water systems.

Across these 17 studies, several key variables—such as water temperature, chlorine levels, and biofilm formation—were repeatedly identified as critical contributors to *Legionella* growth. Machine learning models, as employed in studies like **Xu et al. (2024)** and **Brunello et al. (2022)**, showed great promise in accurately predicting contamination risks, but these models often require extensive validation before they can be broadly applied. Public health interventions, as explored in studies like **Spica et al. (2024)** and **Wilson et al. (2022)**, also play a crucial role in mitigating *Legionella* risks, particularly in high-risk environments like hospitals and agricultural irrigation systems. While the studies agree on the main factors influencing *Legionella* contamination, they differ in their approaches to predicting and mitigating these risks, reflecting the complexity of managing *Legionella* in diverse environmental settings.

Table 5 – Characteristics of the studies.

Study and Country	Objetive	Model	Sample	Design	Evaluation	Limitations
Hancock et al. (2014) England	To develop a mathematical model to predict and control non-transmissible infections, such as Legionnaires' disease, by incorporating human movement data. The study aims to locate infection sources and predict populations at risk during outbreaks.	Human Movement-Based Model for Non-Transmissible Infections.	The study used movement data from the Great Britain Human Movement (GBHM) database and case-report data from three Legionnaires' disease outbreaks in Great Britain (Stoke-on-Trent, Hereford, and Barrow-in-Furness).	The authors applied a spatial interaction model informed by detailed human movement patterns from the GBHM database. The model incorporates socio-demographic data, individual travel behaviour, and infection rates to predict infection source locations. Three levels of data richness (individual travel histories, GBHM database estimates, and basic distance-based movement kernels) were compared for their accuracy in predicting the infection source.	The model accurately identified the source location when travel histories of infected individuals were available (Level 1). However, even without detailed travel data, the GBHM database still provided accurate predictions of the infection source (Level 2). The study demonstrated that using detailed movement data improves the accuracy of predicting the source of Legionnaires' disease outbreaks and the residential locations of infected individuals.	The model depends heavily on the quality and detail of available movement data. In the absence of individual travel histories, predictions become less accurate. The model also assumes a constant infection rate, which may not hold true in all cases, particularly if environmental factors like weather conditions are involved.
Dye. et al (2021) United States of America	To develop a neural network capable of predicting *Legionella* risk in premise plumbing systems based on water quality parameters.	Neural network model developed with the TensorFlow machine learning platform.	The model was trained on 504 water samples, with data on parameters such as total chlorine, free chlorine, temperature, and pH.	A convolutional neural network was built using TensorFlow, with the dataset split into training (362 samples), validation (41 samples), and test sets (101 samples). The network was optimized through several iterations to determine the best-performing architecture.	The model achieved an accuracy of 86.14% on the test set, successfully identifying parameters that were most strongly associated with *Legionella* presence. Chlorine levels and water temperature were found to be the strongest predictors.	The dataset excluded key parameters like turbidity due to insufficient sample sizes. More comprehensive datasets could improve model performance and generalizability.
Brunello et al. (2022) Italy	To use machine learning methods to analyse environmental	The primary model used in this analysis is a tree-based ensemble	About 24,000 water samples were collected	The study used retrospective, unsupervised spatiotemporal data mining and tree-based ensemble	The machine learning models identified several key environmental factors (humidity, temperature, latitude)	The study focuses on environmental surveillance, so direct links to public

Study and Country	Objetive	Model	Sample	Design	Evaluation	Limitations
	Legionella surveillance data from the Friuli-Venezia Giulia region of Italy, covering a period from 2002–2019.	classification method, specifically XGBoostClassifier, which builds decision trees for classification.	during routine environmental surveys.	classification methods to identify patterns of *Legionella* spread. The goal was to predict contamination levels and serogroups based on geographical and climatic factors.	influencing *Legionella* contamination. Spatial clusters of high-risk areas were also identified.	health outcomes were not explored. The complex interactions between environmental factors and *Legionella* strains require further study.
Papagianeli et al. (2021) Greece	To develop and validate a mathematical model for predicting the thermal inactivation of *Legionella pneumophila* in water systems.	Inactivation kinetic model.	Bottled water, commercially available in the Greek market, was selected as a model system in order to assure a constant composition. Water samples were spiked with *L. pneumophila* and subjected to various temperature conditions ranging from 51°C to 61°C.	A predictive microbiology approach was used, where thermal inactivation data were fitted to a primary log-linear model. The model was further validated under dynamic temperature conditions to mimic real-world water systems.	The model accurately predicted *L. pneumophila* inactivation across a range of thermal disinfection scenarios. It showed that higher temperatures (>55°C) resulted in faster pathogen inactivation, while lower temperatures allowed the bacteria to persist.	The study used bottled water as the model system, which may not perfectly replicate real-world water systems with varying chemical compositions. Moreover, the model's performance in complex water distribution systems with biofilms needs further validation.
Girolamini et al. (2023) Italy	To develop a predictive model of bioaerosol dispersion using PM10 particles as a proxy, to generate risk maps of *Legionella* spread in the surrounding area in several weather and	Quasi-Gaussian dispersion model known as the ADMS-Urban model.	The cooling tower (CT) investigated in this study is located on the rooftop of a hospital building located in the center of Bologna, Italy. This CT is an open cooling tower, model VTL-E 245-P, produced by	The physical and chemical parameters of the CT water basin were measured during seven years of the study, and include temperature, pH, hardness, alkalinity, conductivity, and disinfectant residues. A dispersion model using PM_{10} particles as a proxy for bioaerosol	The risk model implemented suggests indeed how the PM_{10} aerosols emitted by the CT disperse over the inhabited area in the surroundings of the CT, notably during the winter season. Indeed, during this season risks of contaminations appear higher, as suggested by the elevated concentrations of PM_{10} aerosols	The study focused on a single cooling tower in a specific geographic area, and additional research is needed to verify the model's accuracy in other contexts.

Study and Country	Objetive	Model	Sample	Design	Evaluation	Limitations
	microbiological conditions.		Baltimore Aircoil Company (BAC) (Jessup, MD, USA), installed in 2013 to serve the air conditioning circuit of the facility.	spread was developed. The study collected water samples from the cooling tower, measuring *Legionella* concentrations and analyzing meteorological data to predict bioaerosol spread up to 1.5 km from the tower.	observed over an extended area relatively distant from the CT.	
Wong et al. (2024) United States of America	To train a deep learning computer vision model to automatically detect cooling towers that are aerially visible and that could present a possible risk.	Two-stage deep learning model combining YOLOv5, an object detection model, and EfficientNet-b5, an image classification model.	The model was trained on 2051 images containing 7292 cooling towers from satellite imagery in New York City and Philadelphia. The test set included images from Boston and Athens, Georgia.	A two-stage model using YOLOv5 and EfficientNet-b5 was developed to detect cooling towers from aerial images. The model's performance was compared against manual identification by epidemiologists.	The model achieved a sensitivity of 95.1% and a positive predictive value (PPV) of 90.1% in New York City, with slightly lower sensitivity (91.6%) and PPV (80.8%) in Boston. It significantly outperformed human investigators in terms of speed, identifying cooling towers 600 times faster.	The model was primarily trained on images from a few cities and may not generalize well to different geographic regions or image qualities. Further validation in other cities and regions would improve its robustness.
De Giglio et al. (2021) Italy	To develop a scoring tool to predict the risk of *Legionella* contamination and Legionnaires' disease outbreaks in touristic and recreational facilities.	Poisson Regression Model.	Data was gathered from 47 facilities in the Apulia region of southern Italy, assessing 18 years of clinical and environmental surveillance, including water samples for *Legionella* testing.	A predictive Poisson regression model was built to calculate a risk score based on 47 structural and management parameters. These factors included facility type, number of floors, water system configuration, and cleaning protocols. Receiver operating characteristic (ROC) curves were used to establish cut-off points for significant contamination and outbreak risk.	The model identified critical thresholds for *Legionella* presence and outbreaks, helping facility managers determine when preventive measures should be implemented. Facilities that exceeded the risk thresholds had a significantly higher likelihood of *Legionella* contamination.	The study's application is regional, and the model's transferability to other geographic or facility types may require adjustments. Additionally, the variability in how facilities manage their water systems could affect the risk predictions.

Study and Country	Objetive	Model	Sample	Design	Evaluation	Limitations
De Giglio et al. (2024) Italy	The study aims to standardize a method for predicting the risk of *Legionella* contamination in the water supply of a hospital facility, by comparing Machine Learning, conventional and combined models.	Machine Learning (ML) Model and Poisson Regression Model.	1,053 water samples collected from hospital rooms over a period of 16 months (July 2021–October 2022). Fifty-eight parameters regarding the structural and environmental characteristics of the water network were collected.	The study used a machine learning approach (with a neural network structure) alongside conventional regression models. The models were built on 70% of the dataset and tested on the remaining 30%. Key metrics evaluated included accuracy, sensitivity, and specificity of the predictions.	The machine learning model demonstrated higher accuracy (93.4%) compared to the regression model (82.9%). Sensitivity (the ability to correctly identify positive samples) was also higher for machine learning (43.8%) versus regression (20.3%). The most critical parameters influencing contamination risk were water type (hot vs. cold), filter replacement, and atmospheric temperature.	The study is limited by its single-hospital sample size and may not generalize to other facilities with different water systems or environmental conditions. The dataset could benefit from including more diverse hospitals to enhance the model's predictive ability.
Chochlakis et al. (2023) Greece	To investigate the effect of physico-chemical and microbiological agents on *Legionella* detection in hotel water distribution systems in Crete, with the aim of developing a more holistic preventive model for *Legionella* outbreaks.	Generalized linear model (GLM) and a binomial model.	Water samples were collected from hotel water systems and tested for *Legionella spp.*, *E. coli*, coliforms, total mesophilic count, and *Pseudomonas aeruginosa*. Other variables like temperature, chlorine levels, number of rooms, and proximity to the boiler were recorded.	This was an observational study using multi-criterion analysis and generalized linear and logistic regression models. The models were built to identify independent predictive factors related to *Legionella* presence.	The study found that higher water temperatures significantly influenced *Legionella* presence, regardless of whether the water was hot or cold. Opportunistic pathogens *like P. aeruginosa* and non-opportunistic ones like *E. coli* were strongly correlated with *Legionella*.	The study did not explore the formation and impact of biofilms in depth, which could be an important factor for *Legionella* survival. Additionally, the hotel's cleaning practices and specific environmental conditions were not fully controlled for.
H. Weir et al. (2019)	To develop a new method for estimating Legionellosis risks in a	Quantitative Microbial Risk Assessment (QMRA) Model.	The study used USA national incidence rates from the Centers for	A computational study using QMRA. The model incorporated demographic variability (age,	The study demonstrated significant differences in health outcomes (severe disability-adjusted life years) between	The model relies on assumptions about exposure volume

Study and Country	Objetive	Model	Sample	Design	Evaluation	Limitations
United States of America	diverse population using Quantitative Microbial Risk Assessment (QMRA). This method addresses within-group heterogeneity, which is often ignored in conventional dose-response models.		Disease Control and Prevention (CDC) as proxies for within-group variability. No direct water samples were used, but the focus was on computational modeling using demographic data.	underlying health conditions, etc.) and considered differences in risk between population subgroups. Data from CDC reports on *Legionella* were used to develop the model, which was then applied to different scenarios to estimate Disability-Adjusted Life Years (DALYs).	demographic groups, particularly in older adults or immunocompromised individuals. The model allowed for more precise risk estimates compared to traditional models.	heterogeneity, and there may be limitations in how well it accounts for individual differences in susceptibility. Additionally, it was not validated with real-time water sample data.
Hamilton et al. (2019) United States of America	To define risk-based critical concentrations of *L. pneumophila* for residential water uses, focusing on indoor fixtures like showers, faucets, and toilets.	Quantitative Microbial Risk Assessment (QMRA) model.	Simulated *Legionella* concentrations in water samples from different indoor residential fixtures. The study used modeling approaches rather than real water samples.	The study used quantitative microbial risk assessment (QMRA) to assess exposure risks across various fixtures. Critical concentrations of *Legionella* for infection were calculated, with different risk targets (infection risk vs disability-adjusted life years, DALY).	Showers posed the highest risk for *Legionella* transmission, with critical concentrations calculated at 14.4 CFU per liter for infection risk and 1.4 × 10^3 CFU per liter for a disability-adjusted life year target. Water-efficient showers posed less risk than conventional ones.	The study is based on simulated data, and there is uncertainty in translating these results to real-world residential settings. Critical concentrations often fall below the detection limit of typical culture-based assays, creating a challenge for practical monitoring.
Massiot et al. (2023) France	To monitor *Legionella* infection risks using a Bayesian network model updated with temporal measurements from agricultural irrigation systems using reclaimed wastewater.	General Bayesian Network (GBN).	Water samples from two corn fields in South-West France, irrigated with wastewater treated by ultra-filtration and UV.	This study used Bayesian networks within a quantitative microbial risk assessment (QMRA) framework. *Legionella* levels were monitored in wastewater before and after treatment. The model was updated with real-time measurements to predict infection	High levels of *Legionella* were detected even after UV treatment, but the median annual infection risk did not exceed safety benchmarks set by the U.S. Environmental Protection Agency (EPA). The Bayesian model successfully adapted to new data to provide updated risk estimates.	The pilot experiment focused on a limited number of irrigation systems, and further validation with different crops and wastewater treatments would enhance the generalizability of the results.

Study and Country	Objetive	Model	Sample	Design	Evaluation	Limitations
				risks in the population living near the fields.		
Van der Ploeg and Steyerberg (2016) Netherlands	To compare different feature selection methods for predicting *L. pneumophila* infection, using a dataset of *Legionella* strains classified as clinical or environmental.	Multiple Models: Classification and Regression Trees (CART), Random Forest (RF), Support Vector Machines (SVM), and Least Absolute Shrinkage and Selection Operator (LASSO).	222 *Legionella pneumophila* strains with 448 continuous DNA markers, classified as either clinical or environmental strains.	The study employed various feature selection methods, including classification and regression trees (CART), random forests (RF), support vector machines (SVM), and least absolute shrinkage and selection operator (LASSO). Each model's ability to discriminate between clinical and environmental strains was assessed using bootstrap sampling.	The random forest and LASSO models performed best, with validated AUC values of 0.975 and 0.925, respectively. These methods provided more reliable predictions compared to traditional statistical methods.	The study is limited to genetic markers, and its focus on feature selection may not directly translate into practical applications for public health interventions.
Wilson et al. (2022) United States of America	To develop a web-based risk calculator that relates *Legionella* water monitoring results (from showers) to health outcomes.	Quantitative Microbial Risk Assessment (QMRA) model.	Simulated water samples with varying concentrations of *Legionella*, analysed using two exposure models.	The study compared two exposure models for estimating infection risks from *Legionella* aerosolization in showers. A Monte Carlo simulation was used to account for variability and uncertainty in exposure parameters.	Model 2, which differentiated between water-efficient and conventional showers, consistently estimated higher infection risks.	The study focused on showers and did not consider other potential *Legionella* sources, such as faucets or cooling towers. Furthermore, the models did not include real-world data.
Ortiz et al. (2024) Canada	To identify critical uncertainty factors in mechanistic models predicting *Legionella* concentrations in building systems.	EPANET-MSX Mechanistic Model.	Data from a monitored water-efficient residence with 58 discrete samples were collected in 2018.	Mechanistic modeling using EPANET-MSX, incorporating nutrient levels, temperature, chlorine, and biofilm interactions.	Biofilm detachment had a significant impact, increasing predicted *Legionella* concentrations by up to 4.2 logs in cold systems.	The model struggled with accurately simulating real-world *Legionella* concentrations due to limited parameter data for nutrient and biofilm dynamics.

Study and Country	**Objetive**	**Model**	**Sample**	**Design**	**Evaluation**	**Limitations**
Spica et al. (2024) Italy	To create a predictive model for evaluating public health strategies to prevent *Legionella* outbreaks in non-hospital buildings.	Mathematical Model for Assessing Prevention Strategies for Legionnaires' Disease.	Data from 243 commercial and residential buildings in Italy were used, focusing on water quality and infrastructure.	Mathematical model using variables such as building type, water temperature, and contamination levels.	The model successfully predicted high-risk buildings and the cost-effectiveness of different prevention strategies.	Limited geographic scope to Italy and a lack of validation against actual outbreak data. Additional data needed for broader applicability.
Xu et al. (2024) United States of America	To develop a machine learning model to predict *Legionella* contamination in premise plumbing systems using water quality data.	Logistic regression and Multivariate Linear Regression (MLR).	192 water samples from a building's plumbing system, measuring concentrations of Legionella, amoebae, and water quality.	Machine learning model using LASSO regression, principal component analysis (PCA), and multivariate linear regression to estimate Legionella concentration.	The model showed good accuracy, with *Legionella* concentration ranges being estimated based on the physiochemical parameters of water.	The model does not differentiate between live and dead *Legionella* organisms, leading to overestimation of contamination.

4.4. Evaluation of Methodology Quality and Study Risk of Bias Assessment

In reviewing the results of the studies using the JBI Critical Appraisal Checklist, the responses to each question reflect the methodological strengths and weaknesses of the included research. The checklist covers key elements critical to evaluating the reliability and validity of the studies, which helps in understanding their contributions to predictive modeling for *Legionella spp*.

Q1: Is the review question clearly and explicitly stated?

This question evaluates whether the study's objective is clear from the outset. For all the studies, the answer was consistently Yes, meaning each study defined its research question and scope clearly, ensuring readers and reviewers understand the study's aim. Having a well-defined question is essential in ensuring that the study remains focused and that the research methodology aligns with the objective, which is crucial for predictive models in complex scenarios like *Legionella* contamination.

Q2: Were the inclusion criteria appropriate for the review question?

The inclusion criteria define what types of data or subjects are acceptable for a study. A consistent Yes across all studies indicates that each study carefully selected relevant data or case studies to address the research question. In the context of *Legionella* predictive modeling, defining appropriate inclusion criteria is critical to ensure that relevant environmental factors, water systems, or case studies are analyzed to provide meaningful predictions.

Q3: Was the search strategy appropriate?

This question is typically more relevant for systematic reviews, where a well-defined search strategy ensures that all relevant studies are included. However, for many of the primary studies in this review, Not applicable was marked since these studies were not reviews of the literature but rather focused on specific datasets or experimental procedures. When applicable, the search strategy ensures that no relevant studies or data sources are overlooked, which is crucial when building comprehensive predictive models.

Q4: Were the sources and resources used to search for studies adequate?

Similarly, Q4 focuses on whether the authors used appropriate and comprehensive sources of data. In all cases, Yes was answered, meaning the studies were pulled from reliable and relevant data sources, such as environmental monitoring data or water quality measurements. Adequate data sources are vital in studies like these because predictive models are only as good as the data they are built on, especially when dealing with a pathogen as complex and environmentally dependent as *Legionella*.

Q5: Were the criteria for appraising studies appropriate?

This question looks at whether the methodology for evaluating the data was sound. For all the studies, Yes was the response, which indicates that the methods for assessing the validity of the data used in predictive models were well-considered. In *Legionella* studies, this could involve assessing water quality, environmental factors, or human movement patterns to determine what data are most relevant to the predictive outcomes.

Q6: Was critical appraisal conducted by two or more reviewers independently?

For most studies, the answer to this was Unclear because it was often not stated whether two or more reviewers evaluated the data. This question ensures that there is reduced bias in data evaluation. In predictive modeling, independent review would minimize bias, ensuring that all relevant data is considered, and no factors are overlooked in the modeling process.

Q7: Were there methods to minimize errors in data extraction?

Here, the answers varied, with some studies showing a clear process for minimizing data extraction errors (Yes), while others either did not specify or lacked clear procedures (No). Minimizing errors in data extraction is critical because errors can lead to incorrect predictions or poor model performance. In the case of *Legionella*, where contamination risk is modeled based on environmental data, even small errors can lead to inaccurate predictions, which can have serious public health implications.

Q8: Were the methods used to combine data appropriate?

In primary studies, Q8 is reinterpreted to assess whether data from different sources or variables were appropriately combined. For studies that combined data from multiple locations, variables (e.g., environmental factors), or sources, the answer was Yes. Combining diverse datasets appropriately is essential in building accurate predictive models, as these models need to account for multiple factors—such as water temperature, microbial content, and chlorine levels—to accurately predict *Legionella* contamination.

Q9: Was the likelihood of publication bias assessed?

For the most part, not applicable was marked because these are primary research studies and not systematic reviews. However, assessing publication bias is typically important in systematic reviews to ensure that studies with negative or inconclusive results are not excluded, as this could skew the overall findings. Although not directly relevant to primary studies, this question highlights the need for transparency in data reporting and comprehensive analysis in predictive modeling.

Q10: Were recommendations for policy and/or practice supported by the reported data?

All studies received a Yes for this question, indicating that their findings and recommendations were backed by the data they presented. In the context of *Legionella* research, this is vital because the models often influence public health policies and water management practices, ensuring that recommendations are well-supported by data is critical for effective decision-making.

Q11: Were there specific directives for new research appropriate?

This question was answered with Yes across all studies, showing that the authors acknowledged the limitations of their research and provided clear recommendations for future studies. In predictive modeling, continuous improvement of models and further research are necessary because environmental conditions and *Legionella* behavior can change over time, making it crucial for researchers to identify areas where further investigation could refine or improve models.

The risk of bias varies among the studies, with some marked as low risk and others as moderate. Studies with low risk demonstrated a strong methodology, clear data extraction processes, and appropriate use of data, making their results more reliable. Studies with moderate risk had minor concerns, such as a lack of clarity on how data extraction errors were minimized or how independent review was conducted, which could introduce some uncertainty into their findings.

The checklist reveals that the studies reviewed for *Legionella* predictive modeling largely meet methodological standards, with some room for improvement, particularly in minimizing errors in data extraction and ensuring independent review. These factors are crucial for ensuring the reliability and accuracy of predictive models, which can have significant public health implications. The results indicate that while the studies provide valuable insights and models, there is a need for consistent data handling procedures to further strengthen the predictive accuracy and applicability of the models.

5. Discussion

This systematic review provides an exploration of predictive models used to assess the risk of *Legionella spp.* contamination and their effectiveness in real-world applications. The review analysed seventeen studies, each employing a variety of models ranging from traditional statistical methods to advanced machine learning techniques. The focus of the review was to identify these predictive models, interpret their mechanisms, evaluate their strengths and weaknesses, highlight the key variables that influence *Legionella* growth, and assess the models' overall importance in predicting and preventing outbreaks.

A wide spectrum of models was identified, reflecting the evolving nature of predictive analytics in managing *Legionella*. Machine learning models, such as those seen in **Xu et al. (2024)** and **Brunello et al. (2022)**, demonstrated substantial accuracy in predicting contamination patterns and estimating *Legionella* concentrations in water systems based on environmental data. **Xu et al. (2024)**, for instance, utilized a neural network model, which leveraged large datasets that incorporated variables like water temperature, chlorine levels, and biofilm dynamics to predict *Legionella* proliferation. Neural networks excel in such applications because of their ability to model complex, non-linear relationships between multiple variables, making them particularly well-suited for dynamic water systems where environmental factors fluctuate.

In contrast, **De Giglio et al. (2021, 2024)** adopted more traditional approaches, using Poisson regression models and risk-scoring tools. These models are simpler in their construction and more easily interpretable, focusing on quantifying contamination risks in controlled environments like hospitals and recreational facilities. The Poisson regression model, for example, provided a straightforward way to model the likelihood of contamination based on risk factors such as water temperature and chlorine concentration. While these models are less adaptable than machine learning algorithms, they offer significant advantages in settings where data is limited or when real-time data processing is unnecessary.

One of the key findings from the analysis of these models is the superior performance of machine learning algorithms, particularly when applied to large and diverse datasets. Models like random forests and neural networks—used in **Brunello et al. (2022)** and **Xu et al. (2024)** — consistently outperformed traditional methods in terms of accuracy and predictive power. For example, **De Giglio et al. (2024)** demonstrated that their machine learning model achieved 93.4% accuracy, significantly higher than the 82.9% accuracy of their conventional Poisson regression model. Machine learning models are particularly valuable because of their ability to integrate a wide variety of variables, such as temperature, biofilm presence, and chlorine concentration, and dynamically adjust to new inputs. This allows for real-time monitoring and predictions, a critical capability in high-risk environments like healthcare facilities and industrial water systems.

However, while machine learning models like those in **Wong et al. (2024)** – which utilized deep learning to detect cooling towers – exhibit high sensitivity (over 95% in detecting potential contamination sources), they are not without limitations. One significant challenge is the dependency on large, high-quality datasets. Machine learning models require extensive training data to function effectively, and their accuracy can suffer when faced with incomplete or inconsistent data. For instance, **Dye et al. (2021)** excluded parameters like turbidity due to insufficient data, which hindered the model's ability to generalize across different water systems. Similarly, **Xu et al. (2024)**'s model was limited to a single building and did not account for the differentiation between live and dead *Legionella* organisms, potentially overestimating contamination risks.

Traditional statistical models, while less accurate in some cases, still have their place in predictive modeling, especially in settings with fewer resources or limited data availability. These models, like the Poisson regression model in **De Giglio et al. (2021)**, are easier to implement and interpret, making them useful in smaller-scale applications where simplicity and transparency are prioritized over accuracy. Moreover, these models are beneficial in environments where the complexity of machine learning algorithms may pose challenges for public health officials and facility managers. In these contexts, a simple, interpretable model that can offer actionable insights is often more useful than a highly accurate but opaque machine learning model.

A commonality among all the models reviewed is the emphasis on key environmental variables that influence *Legionella* growth. Across the studies, water temperature emerged as one of the most significant factors. *Legionella* bacteria thrive in temperatures between 25°C and 45°C, and models like those in **Brunello et al. (2022)** and **De Giglio et al. (2024)** identified temperature as a major predictor of contamination risk.

In addition to temperature, chlorine levels were consistently identified as another critical variable. Chlorine is widely used to disinfect water systems, but its efficacy depends on maintaining sufficient concentrations to inhibit *Legionella* growth. Models like those in **Dye et al. (2021)** demonstrated that lower chlorine levels were strongly correlated with increased *Legionella* contamination, underscoring the importance of continuous monitoring of disinfectant levels in water systems. **Ortiz et al. (2024)** further emphasized the role of biofilms in promoting *Legionella* proliferation, noting that biofilm detachment events could significantly increase contamination levels, particularly in systems with inadequate chlorine levels.

Environmental factors, such as humidity and precipitation patterns, also play a vital role in influencing *Legionella* growth. For example, **Girolamini et al. (2023)** showed that particulate matter (PM_{10}) and meteorological conditions had a significant impact on the spread of *Legionella* from cooling towers. These findings highlight the need for predictive models to integrate not only microbiological factors but also broader environmental conditions, particularly as climate change alters weather patterns and increases the prevalence of favourable conditions for *Legionella* growth.

The review also highlighted the importance of assessing the public health impact of predictive models. Models such as those in **H. Weir et al. (2019)** and **Hamilton et al. (2019)** used Quantitative Microbial Risk Assessment (QMRA) frameworks to estimate the likelihood of infection in different exposure scenarios. These models were particularly effective in quantifying the risk to vulnerable populations, such as older adults and immunocompromised individuals, who are more susceptible to *Legionella* infections. The ability to predict infection risks in specific populations is critical for informing public health policies, particularly in high-risk environments like hospitals, where proactive measures can prevent outbreaks.

A crucial aspect of these models is their capacity to inform public health strategies. In healthcare settings, for example, **De Giglio et al. (2024)** demonstrated how predictive models could be used to optimize water system maintenance schedules, ensuring timely filter replacements and reducing the risk of contamination. These practical applications highlight the potential of predictive models to go beyond risk assessment and play an active role in outbreak prevention.

Another significant factor to consider is the impact of climate change on *Legionella* contamination risks. As global temperatures rise and weather patterns shift, the conditions that favour *Legionella* growth — such as warmer temperatures, increased humidity, and stagnant water—are becoming more common. Studies like **Brunello et al. (2022)** and **Girolamini et al. (2023)** noted the role of microclimatic conditions in influencing contamination patterns. Additionally, **Massiot et al. (2023)** highlighted the risks posed by water recycling systems in agricultural settings, noting that even after wastewater treatment, *Legionella* levels could remain high, posing a risk of aerosolization and infection, particularly in regions experiencing water shortages or droughts.

As these environmental conditions continue to evolve, it is crucial that predictive models adapt to account for the new challenges posed by climate change. Models must integrate climate-related variables, such as temperature increases and changing precipitation patterns, to ensure accurate risk assessments in the face of these emerging threats. Furthermore, as water systems become more vulnerable to contamination, particularly in the context of extreme weather events, predictive models will need to incorporate more real-time monitoring capabilities and become more versatile in their applications.

In conclusion, the predictive models analysed in this systematic review offer valuable insights into the factors that contribute to *Legionella* growth and infection risk. While machine learning models have demonstrated superior accuracy and adaptability, their complexity and reliance on large datasets present challenges for practical application. Traditional statistical models, while less precise, offer a simpler and more interpretable solution, particularly in settings with limited data. The key variables identified—water temperature, chlorine levels, and biofilm presence—are critical for understanding and mitigating *Legionella* contamination. However, these must be contextualized within the broader environmental changes driven by climate change, which will require predictive models to evolve and incorporate new factors to remain relevant and effective in safeguarding public health.

6. Conclusion

This systematic review set out to evaluate predictive models for *Legionella spp.* infections and assess their reliability for real-world prevention efforts. The analysis of the included studies highlighted the potential of predictive modeling to significantly enhance public health interventions by forecasting *Legionella* outbreaks and informing risk management.

A key finding across the studies is the importance of accounting for environmental, biological, and infrastructural factors that affect *Legionella* growth. Water temperature, chlorine levels, biofilm formation, and environmental conditions like humidity and particulate matter were consistently identified as critical variables. Warmer temperatures and stagnant water conditions, coupled with inadequate disinfectant levels, provide ideal conditions for *Legionella* to proliferate. This highlights the necessity for models to integrate these variables comprehensively to produce accurate risk assessments.

Machine learning models, particularly neural networks and random forests have emerged as the most promising tools due to their ability to process large datasets and dynamically adjust to new information. These models demonstrated higher predictive accuracy than traditional statistical approaches, offering timely insights that can aid in public health interventions. However, their complexity and lack of transparency present challenges, particularly in terms of interpretability for non-specialists. On the other hand, traditional statistical models provide greater clarity and ease of use, making them more applicable in controlled environments, though they often lack the precision needed to account for complex interactions between variables.

The review also identified certain limitations, particularly regarding the generalizability of many models. Several studies relied on localized datasets, which limits the applicability of their findings to broader geographic and environmental contexts. The models' effectiveness in other regions, particularly in areas with different climates or water system infrastructures, remains uncertain. This highlights the need for future models to incorporate more diverse datasets to enhance their applicability across various environments.

Another critical observation is the role of biofilms in *Legionella* growth. Biofilms provide a protective environment for the bacteria, making it difficult for standard water treatment methods, such as chlorination, to fully eradicate the risk. Models that do not account for biofilm dynamics may underestimate contamination risks. Moving forward, predictive models will need to incorporate real-time biofilm monitoring to improve accuracy.

The review also addressed the impact of climate change on *Legionella* proliferation. Rising temperatures and changing precipitation patterns are likely to create conditions more conducive to bacterial growth, increasing the need for models to incorporate climate-related factors. Water systems in regions prone to higher temperatures and increased rainfall will face greater challenges in managing *Legionella*, making it imperative that predictive models evolve to address these emerging risks.

Finally, predictive models are invaluable in shaping public health strategies by identifying high-risk areas and enabling preventive measures before outbreaks occur. While no single model provides a universal solution, the combined use of machine learning and statistical approaches offers a promising path forward. By leveraging the strengths of both, predictive models can become more accurate, adaptable, and accessible, ultimately helping to reduce the incidence of *Legionella*.

In conclusion, the models reviewed in this systematic review demonstrate significant potential in preventing *Legionella* outbreaks, but they require continuous refinement. The integration of more diverse datasets, consideration of biofilm dynamics, and adaptation to climate change will ensure that predictive models become reliable tools in safeguarding public health. With ongoing advancements, these models can be optimized to meet the challenges of increasingly complex water systems and environmental conditions, making them essential components in managing *Legionella* risks in the future.

7. Limitations of the Study

This systematic review has some limitations tied to its design. First, the focus on studies from 2014 to 2024, and the restriction to English, Portuguese, and Spanish publications, may have excluded relevant research. The use of specific databases (PubMed, ScienceDirect, and Web of Science) might have limited the scope, potentially missing significant works published elsewhere. Additionally, while the SPIDER framework effectively defined eligibility, it might have filtered out studies with different but valuable approaches. Variability across the studies in terms of methodology, model types, and environments further complicates direct comparison and generalization of results. Finally, the heterogeneity of the data and lack of full methodological details in some studies made consistent quality assessment more difficult.

8. Suggestions for Future Works

Future research should aim for a broader data collection scope, incorporating studies from underrepresented regions and expanding the geographic relevance of predictive models. Enhancing the interpretability of machine learning models, particularly for public health officials, is crucial. Combining the precision of machine learning with the transparency of traditional models may provide more effective solutions. Integrating real-time monitoring technologies and incorporating climate change factors into predictive models will further improve their accuracy and adaptability, ensuring they remain relevant as environmental conditions shift.

References

CDC. (September de 2024). *How Legionella Spreads*. Obtido de U.S. Centers for Disease Control and Prevention: https://www.cdc.gov/legionella/causes/index.html

Circular Normativa nº 6/DT. (22 de 04 de 2004). Direção-Geral da Saúde (DGS). Obtido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://www.arsalentejo.min-saude.pt/utentes/saudepublica/AreasSaude/DoencasInfecciosas/Documents/CNormativa%20n%C2%BA%206_DT.pdf

Decreto-Lei n.º 69/2023. (21 de Agosto de 2023). Diário da República n.º 160/2023, Série I de 2023-08-21. Obtido de https://diariodarepublica.pt/dr/detalhe/decreto-lei/69-2023-220113533

Directive (EU) 2020/2184 of the European Parliament and of the Council of 16 December 2020 on the quality of water intended for human consumption. (16 de December de 2020). *435*, 1-62. Official Journal of the European Union. Obtido de https://eur-lex.europa.eu/eli/dir/2020/2184/oj

ECDC. (2023). *Legionnaires' disease. Annual Epidemiological Report for 2021.* Stockholm: European Centre for Disease Prevention and Control.

Edelstein, P. (2007). Urine antigen tests positive for Pontiac fever: implications for diagnosis and pathogenesis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 44*(2), 229-231. doi:https://doi.org/10.1086/510394

Kanarek, P., Bogiel, T., & Breza-Boruta, B. (2022). Legionellosis risk-an overview of *Legionella spp*. habitats in Europe. *Environmental science and pollution research international, 29*(51), 76532-76542. doi:https://doi.org/10.1007/s11356-022-22950-9

Lei n.º 52/2018 da Assembleia da República. (20 de Agosto de 2018). 4229 - 4251. Diário da República: n.º 159/2018, Série I. Obtido de https://data.dre.pt/eli/lei/102/2009/09/10/p/dre/pt/html

Miyashita, N., Higa, F., Aoki, Y., Kikuchi, T., Seki, M., Tateda, K., . . . Watanabe, A. (2020). Distribution of *Legionella* species and serogroups in patients with cultureconfirmed *Legionella* pneumonia. *Journal of Infection and Chemotherapy, 26*(5), 411-417. doi:https://doi.org/10.1016/j.jiac.2019.12.016

Mraz, A. L., & Weir, M. H. (2018). Knowledge to Predict Pathogens: *Legionella pneumophila* Lifecycle Critical Review Part I Uptake into Host Cells. *Water, 2018, 10,132*. doi:doi:10.3390/w10020132

NASEM. (2020). *Management of Legionella in Water Systems.* Washington,DC: The National Academies Press. doi:https://doi.org/10.17226/25474

Norma n.º 024/2017. (23 de Junho de 2017). Direção-Geral da Saúde (DGS). Obtido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.lexpoint.pt/Fileget.aspx?FileId=25457

Orientação n.º 20/2017. (25 de Setembro de 2017). Direção-Geral da Saúde (DGS). Obtido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.sip-spp.pt/media/33wky1i1/legionella-diagno-stico-laboratotial-2017-dgs.pdf

Page, M. J., McKenzie, J. E., Bossuyt, P. M., I., B., Hoffmann, T. C., & M. C. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology, 134*, 178-189.

Sharma, M., Mishra, P., Ali, A., Kumar, P., Kapil, P., Grover, R., . . . Kulshrestha, S. (2023). Microbial Waterborne Diseases in India: Status, Interventions, and Future Perspectives. *Current microbiology, 80*(12), 400. doi:https://doi.org/10.1007/s00284-023-03462-2

Shimada, S., Nakai, R., Aoki, K., Shimoeda, N., Ohno, G., Kudoh, S., . . . Tateda, K. (2021). Chasing Waterborne Pathogens in Antarctic Human-Made and Natural Environments, with Special Reference to *Legionella spp. Applied and Environmental Microbiology, 87.* doi:https://doi.org/10.1128/AEM.02247-20

Tan, L. T., Tee, W. Y., Khan, T. M., Ming, L. C., & Letchumanan, V. (2021). *Legionella pneumophila* — The causative agent of Legionnaires' disease. *Progress in Microbes and Molecular Biology, 4*(1). doi:10.36877/pmmb.a0000193

UN. (2023). *The Sustainable Development Goals Report 2023.* New York: United Nations Publications issued by the Department of Economic and Social Affairs (DESA).

UN. (2024). *World Population Prospects 2024: Summary of Results.* New York: United Nations Department of Economic and Social Affairs, Population Division.

Walker, J. T. (2018). The influence of climate change on waterborne disease and *Legionella*: a review. *Perspectives in Public Health, 138*(5), 282-286. doi:10.1177/1757913918791198

Whiting, P., Rutjes, A., Reitsma, J., Bossuyt, P., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *3*(25). doi:doi:10.1186/1471-2288-3-25

Annexes

Annex A

JBI CRITICAL APPRAISAL CHECKLIST FOR SYSTEMATIC REVIEWS AND RESEARCH SYNTHESES

Reviewer______________________________________ Date______________________________

Author______________________________________ Year_________ Record Number_________

		Yes	No	Unclear	Not applicable
1.	Is the review question clearly and explicitly stated?	☐	☐	☐	☐
2.	Were the inclusion criteria appropriate for the review question?	☐	☐	☐	☐
3.	Was the search strategy appropriate?	☐	☐	☐	☐
4.	Were the sources and resources used to search for studies adequate?	☐	☐	☐	☐
5.	Were the criteria for appraising studies appropriate?	☐	☐	☐	☐
6.	Was critical appraisal conducted by two or more reviewers independently?	☐	☐	☐	☐
7.	Were there methods to minimize errors in data extraction?	☐	☐	☐	☐
8.	Were the methods used to combine studies appropriate?	☐	☐	☐	☐
9.	Was the likelihood of publication bias assessed?	☐	☐	☐	☐
10.	Were recommendations for policy and/or practice supported by the reported data?	☐	☐	☐	☐
11.	Were the specific directives for new research appropriate?	☐	☐	☐	☐

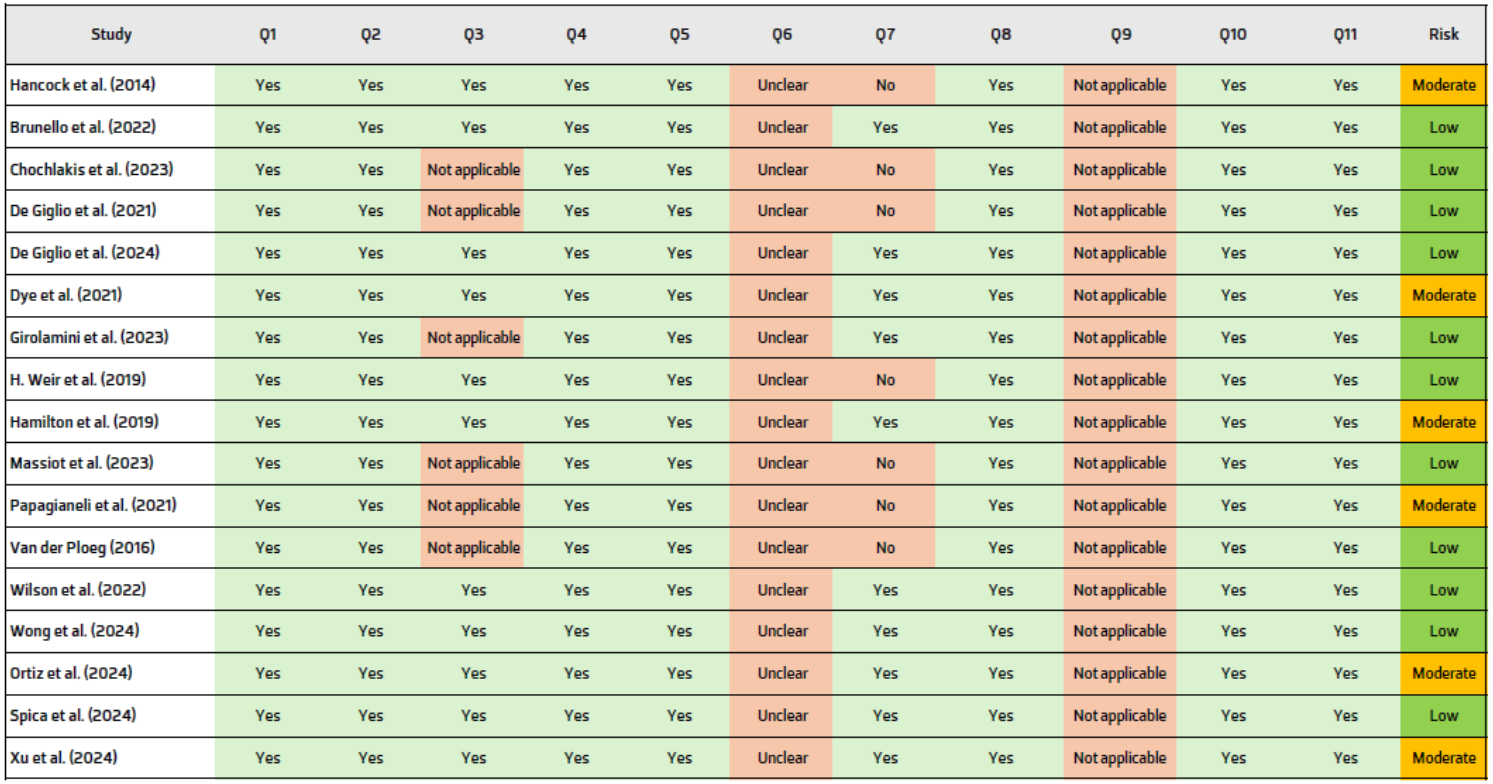
Overall appraisal: Include ☐ Exclude ☐ Seek further info ☐

Comments (Including reason for exclusion)

Critical Appraisal Checklist for Systematic Reviews and Research Syntheses - 3

Annex B

Table 6 – Evaluation of Methodology Quality and Study Risk of Bias Assessment . Based on (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003).



Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Risk
Hancock et al. (2014)	Yes	Yes	Yes	Yes	Yes	Unclear	No	Yes	Not applicable	Yes	Yes	Moderate
Brunello et al. (2022)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Low
Chochlakis et al. (2023)	Yes	Yes	Not applicable	Yes	Yes	Unclear	No	Yes	Not applicable	Yes	Yes	Low
De Giglio et al. (2021)	Yes	Yes	Not applicable	Yes	Yes	Unclear	No	Yes	Not applicable	Yes	Yes	Low
De Giglio et al. (2024)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Low
Dye et al. (2021)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Moderate
Girolamini et al. (2023)	Yes	Yes	Not applicable	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Low
H. Weir et al. (2019)	Yes	Yes	Yes	Yes	Yes	Unclear	No	Yes	Not applicable	Yes	Yes	Low
Hamilton et al. (2019)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Moderate
Massiot et al. (2023)	Yes	Yes	Not applicable	Yes	Yes	Unclear	No	Yes	Not applicable	Yes	Yes	Low
Papagianeli et al. (2021)	Yes	Yes	Not applicable	Yes	Yes	Unclear	No	Yes	Not applicable	Yes	Yes	Moderate
Van der Ploeg (2016)	Yes	Yes	Not applicable	Yes	Yes	Unclear	No	Yes	Not applicable	Yes	Yes	Low
Wilson et al. (2022)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Low
Wong et al. (2024)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Low
Ortiz et al. (2024)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Moderate
Spica et al. (2024)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Low
Xu et al. (2024)	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Not applicable	Yes	Yes	Moderate

1st step – Counting positive and negative answers and analysis of specific criteria. Positive answers have a green background (Yes) and negative answers have an orange background (No, Not applicable and Unclear).

2nd step – Bias Risk Classification.

Low	Moderate	High