World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

ENVIRONMENTAL HEALTH SURVEILLANCE AND PREDICTIVE RISK ASSESSMENT: A SYSTEMATIC REVIEW OF DATA-DRIVEN APPROACHES FOR DETECTING RECURRING HAZARDS IN RESIDENTIAL AND INSTITUTIONAL BUILDINGS

Abhishek Benke *

Students Landing Inc,
* Corresponding Author

ORCID Details
Abhishek Benke: https://orcid.org/0009-0001-7277-4131

World Journal of Advanced Engineering Technology and Sciences, 2026, 21(02), 164–176

Publication history: Received on 06 September 2026; revised on 02 November 2026; accepted on 04 November 2026

Article DOI: https://doi.org/10.30574/wjaets.2026.21.2.0473

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Background: Lead paint, pests, dampness, Legionella in water systems, carbon monoxide, radon and indoor heat seldom affect a building just once. They repeat at the same addresses, building type, and at the same owners' portfolios. Surveillance is still largely reactive; that is, cases are recorded once they are exposed and inspections are carried out in response to complaints. There is now administrative, sensor and digital-trace data to detect problems earlier and proactively.

Objective: To consider peer-reviewed research on assessing, forecasting and/or identifying reoccurring environmental hazards in residential and institutional buildings using data-driven approaches and describe their data, methods, validation, and equity implications.

Methods: Backward and forward citation chaining was performed to conduct a search of bibliographic databases and scholarly web indexes. Studies that were judged to be appropriate were quantitative in nature, and were used to detect, map, predict, or prioritize a hazard at the building, dwelling, or small-area level. Comparators for inspection targeting included studies involving food premises. Narrative synthesis was then carried out for the hazard, setting, data, method, and evidence level.

Results: Twenty-two studies between 2009 and 2024 were analyzed for nine hazards and five method families, with the most prevalent ones being risk prediction and inspection. In the most advanced model (lead poisoning), models that used blood-lead levels, inspection information, and building records performed better than current screening methods, and one validated model improved the positive predictive value of the previous models. Other instances, the true positive rates were as high as 71% for fires, and the number of findings for unsafe food premises were doubled. The most effective variables in capturing recurrence were address and owner history, cooling-tower registration, and the density of code violations. But only five of the studies assessed have reported the ability to place the methods into practice, and complaint information was not comprehensive enough to capture some neighborhoods.

Conclusion: To make building-hazard management preventive, apply data-driven surveillance, including cross-agency data linkage and prospective evaluation audited for equity.

Keywords: Environmental health surveillance, Predictive risk assessment, Machine learning, Housing, Lead poisoning, Legionella, Indoor environment, Institutional buildings, Inspection targeting, Public Health informatics, Systematic review.

1 INTRODUCTION

Public health surveillance involves the continuous and systematic collection, analysis, and interpretation of health data in order that public health practice may be planned, implemented, and evaluated [1]. In the case of environmental hazards, things become more complicated because the relevant information is kept by a number of different agencies—the health departments possess the case reports and laboratory results, the housing and building departments have the inspection and code-enforcement records, and the environmental agencies carry out the measurements. The U.S. National Environmental Public Health Tracking Program was set up specifically to remove this gap in information by combining data on hazards, exposure, and health outcomes [2].

It is in buildings that a great deal of these connections become important. Substandard housing has been linked to a wide variety of infectious, chronic, and injury-related outcomes [3], and data from national surveys have demonstrated that the hazards caused by lead-based paint were still found in a large number of the houses in the United States many years after the use of lead paint in homes had been banned [4]. A key characteristic of many building-related hazards is that they tend to recur. The physical cause—such as deteriorating paint, a leaking building envelope, an unmaintained cooling tower, or an unvented appliance—stays in place until it is dealt with, and similar conditions are likely to reappear in buildings that are owned by the same person, are the same age, and are maintained in the same way. For instance, clinicians in Cincinnati found a group of children who had asthma, developmental difficulties, and high levels of lead living in housing units that were part of a single owner's portfolio [5], and the density of housing-code violations at the neighborhood level has been connected to visits to emergency departments and hospitals for childhood asthma [6]. In New York City, repeated community outbreaks of Legionnaires' disease that were related to building cooling towers resulted in the enactment of laws which required the towers to be registered, inspected, and tested regularly [7].

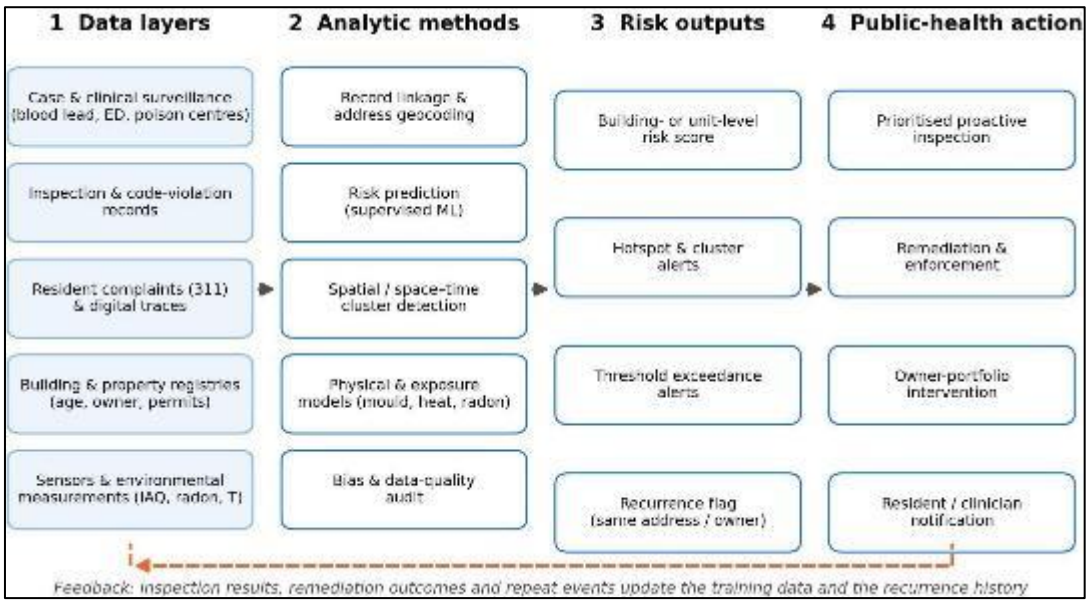


Figure 1: The conceptual framework for the data-driven surveillance of recurring building hazards is such that the various data layers are linked at the address or building level and are then analysed using a range of complementary methods in order to generate risk outputs which lead to public-health action.

Traditional environmental health surveillance is basically a reactive system since a child has to have a high level of lead in their blood or a resident has to make a complaint before a property is checked. On the other hand, data-driven methods take a different approach in that they use administrative, sensor, and digital-trace data to estimate which buildings are most likely to contain a hazard before any real damage takes place [8]. This review looks at three questions: (i) which data-driven techniques have been used to detect, predict, or prioritise recurring environmental hazards in residential and institutional buildings; (ii) what data sources do they make use of and how well have they

been validated; and (iii) what limitations—especially those relating to data bias and equity—prevent their use. Figure 1 shows the conceptual framework which structures the synthesis.

2 METHODS

2.1 Review Design and Reporting

The study was carried out by means of a systematic methodology and has been presented in line with the relevant sections of the PRISMA 2020 statement [9]. The protocol was not registered in advance; since the literature covers areas such as public health, environmental science, building science and computer science, and since many relevant studies are published in conference proceedings and not in indexed journals, the search included both database and web-scale index searches together with citation chaining rather than relying on a single database.

2.2 Search Strategy

The terms in the three concept blocks were searched: (a) hazard terms (lead, lead service line, mold, dampness, pests, rodents, Legionella, carbon monoxide, radon, indoor heat, fire, housing, housing conditions, housing issues, housing quality, code violations); (b) setting terms (housing, residential, dwelling, building, school, hospital, health care facility, nursing home, institutional); (c) method terms (surveillance, prediction, predictive model, machine learning, risk score, spatial analysis, cluster detection, sensor, monitoring, inspection targeting).The search was conducted in bibliographic databases and scholarly web indexes, and followed by checking the reference lists and the articles that cited the studies identified in the search for any additional papers up to 2024.

2.3 Eligibility Criteria

A study was included only if it (i) dealt with an environmental hazard that occurred in or arose from residential or institutional buildings or from their engineered systems; (ii) used a data-driven, quantitative approach—such as statistical or machine-learning prediction, spatial or spatio-temporal analysis, or sensor-based methods. The assessment was carried out using modelling or by means of structured, multi-source surveillance at the level of buildings, dwellings, or small areas, and also considered the reported data sources together with the quantitative or clearly described results. Studies involving commercial food premises were included as methodological comparisons since they offer the most thorough evidence regarding predictive inspection targeting, a feature which can be directly applied to inspections of housing and institutional premises. Studies that were based entirely on building physics without making any use of surveillance or risk assessment, those that focused exclusively on occupational health, and opinion articles were excluded. Although methodological and review papers were used to interpret the studies that were included, these papers were not counted as part of the studies themselves.

2.4 Data Extraction and Evidence Classification

For all the studies listed, the hazard domain, the type of building setting (whether residential, institutional or commercial), the main data sources, the analytic method employed, and the key findings that were reported were all obtained. The studies were then divided into one of five method categories (as shown in Figure 1): risk prediction and inspection targeting; spatial and cluster analysis; measurement and exposure modelling; case- and event-driven surveillance and registries; and data-quality and equity audit. Since the designs of the studies included are too diverse to allow the use of a single risk-of-bias tool, they were classified according to their level of evidence: *descriptive or associational* (in this case the method describes or explains the distribution of the hazard without having been validated predictively); *retrospective validation* (in which the predictive performance was assessed using data that had been set aside or collected at a later date); or *operational deployment* (in which the method was used to guide actual inspections, carry out remediation, or regulate, with the results being reported). The attributes suggested for evaluating public health surveillance systems—namely sensitivity, predictive value positive, representativeness, timeliness, and acceptability—were used to organize the appraisal [10].

2.5 Synthesis

A meta-analysis was not carried out since the hazards, the outcomes, and the performance criteria were so different; the findings were instead combined in a narrative way by method family and were checked in relation to both the hazard domain (see Figure 2) and the publication year (see Figure 3).

3 RESULTS

3.1 Characteristics of the Included Studies

There were 22 studies that met the inclusion criteria (Table 1). They were conducted from 2009 to 2024 and dealt with nine hazard areas: lead (n = 5), food-premises hygiene (n = 3, methodological comparators), pests and general housing conditions (n = 3), Legionella (n = 3), radon (n = 2), heat (n = 2), complaint-data bias (n = 2), fire (n = 1), and carbon monoxide (n = 1). Fourteen studies were conducted in residential buildings, five in institutional buildings or building water systems, and three in commercial food premises (including a study conducted by the fire inspector that included commercial and institutional buildings). The most common method groups included were risk prediction and inspection targeting (n = 9), spatial and cluster analysis (n = 4), case- and event-driven surveillance (n = 4), measurement and exposure modelling (n = 3) and data-quality and equity audit (n = 2). The spread of the evidence by hazard/ method combinations is illustrated in Figure 2: the lead studies all involved risk prediction, the Legionella studies all followed event-driven surveillance and most of the cells in the matrix are blank.

Table 1: Characteristics of the 22 included studies, grouped by method family.

Study	Hazard; setting	Principal data sources	Method	Evidence level	Key reported finding
Risk prediction & inspection targeting					
Kang et al. (2013) [11]	Food hygiene; commercial	Online reviews linked to inspection records	Text classification	Retrospective	Review text predicted hygiene violations beyond inspection history alone
Potash et al. (2015) [8]	Lead; residential	~2.5 M blood-lead tests, ~120,000 home inspections, building and census data	Supervised classification	Retrospective	Address-level features (building age, condition, prior poisoning and inspections) most predictive; precision exceeded baseline at all intervention levels
Madaio et al. (2016) [12]	Fire; commercial and institutional	Fire incidents, inspections, property and business records	Supervised ML	Deployed	True-positive rate up to 71% across >5,000 buildings; 6,096 uninspected properties identified
Schomberg et al. (2016) [13]	Food hygiene; commercial	Online reviews, inspection records (two U.S. cities)	Text-based prediction	Retrospective	Social-media signals supplemented routine inspection targeting
Abernethy et al. (2018) [14]	Lead service lines; residential	Municipal parcel records, service-line excavation results	Adaptive statistical model / active learning	Deployed	Guided home selection for lead pipe excavation under incomplete records
Sadilek et al. (2018) [15]	Foodborne illness; commercial	Aggregated search and location logs	ML classifier (FINDER)	Deployed	52.3% of flagged premises deemed unsafe vs 24.7% under baseline inspection
Potash et al. (2020) [16]	Lead; residential	2.5 M blood tests, 70,000 investigations, 2 M permits and violations	Supervised ML	Retrospective	Positive predictive value 15.5%, about twice that of earlier models; comparable performance across racial and ethnic groups
Lobo et al. (2021) [17]	Lead; residential (area level)	Socio-economic, housing, water-quality data	Random forest	Retrospective	ROC AUC 0.91 (Massachusetts) and 0.85 (New York)

Lobo et al. (2022) [18]	Lead in drinking water; schools	Public water-chemistry, census, and geographic data for >8,000 schools	Supervised ML	Retrospective	High-risk schools identifiable from publicly available data alone
Spatial & cluster analysis					
Reid et al. (2009) [19]	Heat; residential (community)	Census, land cover, air-conditioning prevalence	Factor analysis; vulnerability index	Descriptive	Mapped community determinants of heat vulnerability
Beck et al. (2012) [5]	Multiple housing hazards; residential	Clinical records, housing and ownership records	Case clustering and medical-legal intervention	Deployed	45 children in 16 units of one owner's portfolio; systemic repairs across 19 building complexes
Beck et al. (2014) [6]	Housing code violations; residential	Code-violation and hospital-utilisation data	Spatial ecological analysis	Descriptive	Violation density associated with paediatric asthma ED and hospital use
Walsh (2014) [20]	Rodents; residential	311 rat sightings, housing and census data	Spatial regression; kernel density	Descriptive	Sightings associated with housing characteristics, sociodemographics and open space
Measurement & exposure modelling					
White-Newsome et al. (2012) [21]	Indoor heat; residential (older adults)	Hourly indoor temperature in 30 homes; weather data	Mixed linear regression	Descriptive	Outdoor temperature, solar radiation and dew point explained 38% of indoor temperature variability
Kropat et al. (2015) [22]	Radon; residential	National indoor radon measurement database	Kernel regression mapping	Retrospective	Predictive mapping of indoor radon in complex terrain
Li et al. (2021) [23]	Radon; residential	Indoor radon tests, geological and meteorological data	Ensemble learning	Retrospective	Monthly community-level radon prediction for the greater Boston area
Case- & event-driven surveillance, registries					
Iqbal et al. (2012) [24]	Carbon monoxide; residential	Mortality, ED, hospital and poison-centre data	Multi-source surveillance framework	Descriptive	National framework and burden estimates for CO poisoning
Garrison et al. (2016) [25]	*Legionella*; building water systems	Outbreak investigation reports, 2000–2014	Structured review of investigations	Descriptive	Environmental-control deficiencies in water management identified as the common preventable driver
Fitzhenry et al. (2017) [7]	*Legionella*; cooling towers	Case surveillance, molecular typing, tower registry	Outbreak surveillance and registry	Deployed	6 outbreaks, 213 cases, 18 deaths (2006–2015); 3,909 towers inspected after 2015 law, 46 above action thresholds

Soda et al. (2017) [26]	*Legionella*; health care facilities	Case surveillance, 20 states and one metropolitan area	Surveillance data analysis	Descriptive	Health-care-associated cases identified across many facilities, supporting water-management programmes
Data-quality & equity audit					
Kontokosta and Hong (2021) [27]	Complaint-data bias; residential	311 service requests, building and census data	Spatial modelling of complaint propensity	Descriptive	Propensity to complain varies socio-spatially, biasing complaint-driven decisions
Liu et al. (2024) [28]	Complaint-data bias; residential	311 incident reports (New York City, Chicago)	Statistical estimation of reporting delay	Retrospective	Substantial spatial and socio-economic disparities in reporting delay

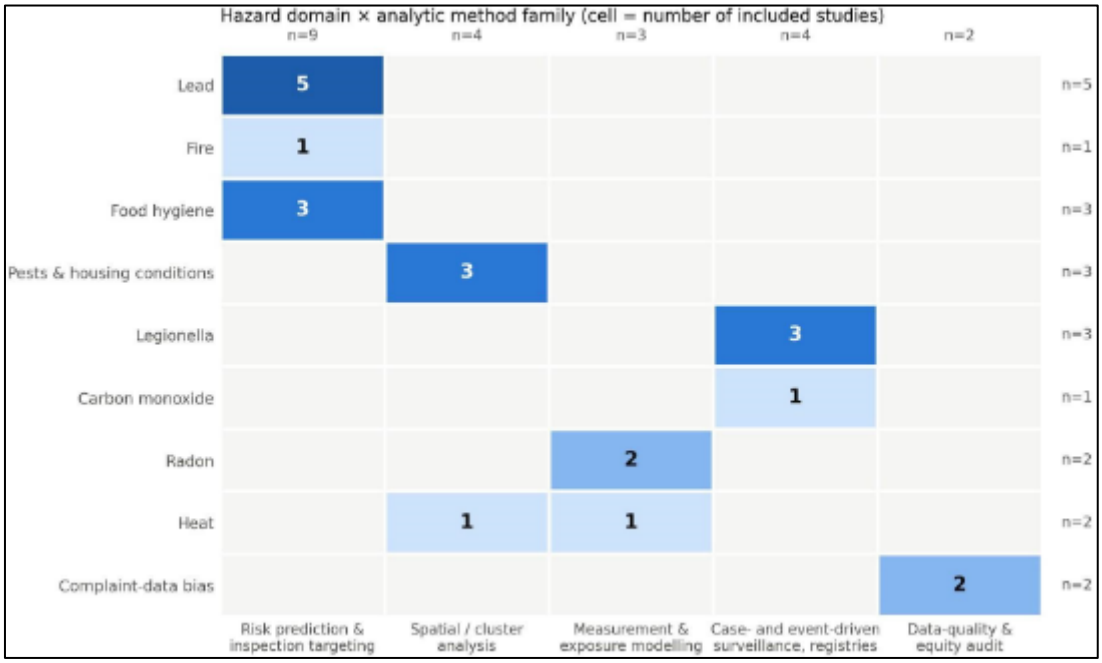


Figure 2: A map showing the evidence from the 22 included studies, organized by hazard domain and analytic method family. In each cell the number of studies is given, with darker shading representing a greater number of studies. The totals for the rows and columns are provided at the sides. The sparse, block-diagonal pattern shows that most hazards have been investigated using only one family of methods.

3.2 Data Sources

The five main types of data used in the studies in question are summarised in Table 2 along with their principal advantages and biases. Case and clinical surveillance data (for example, blood-lead tests, visits to emergency departments, calls to the poison centre and reports of notifiable diseases) serve as the outcome labels against which predictive models are trained, yet they only capture harm that occurs after exposure and rely on who in fact gets tested [16], [24]. Inspection records and the documentation of code violations are the most complete source of hazard information at the level of individual buildings and give the best direct evidence of recurrence, but they are affected by past enforcement priorities [8], [6]. Complaints from residents, particularly those submitted through the municipal 311 system, have the benefit of being both numerous and up-to-date, but they systematically fail to cover all neighborhoods [27], [28]. Data from building and property registers—such as the year of construction, ownership information, permits and the registration of high-risk equipment—supply the stable structural features that make buildings predictable and allow hazards to be linked to owner portfolios [5], [7]. Finally, environmental measurements obtained from monitoring programmes and from low-cost sensor networks provide direct information on exposure but are still limited in quantity, and the low-cost devices have to be carefully calibrated [29], [30].

Table 2: Data sources used for data-driven surveillance of building hazards, with principal strengths and limitations

Data source	Examples in included studies	Hazards addressed	Strengths	Principal limitations
Case and clinical surveillance	Blood-lead tests; ED and hospital records; poison-centre calls; notifiable disease reports [16], [24], [26]	Lead, CO, *Legionella*, asthma triggers	Health-relevant outcome labels; long time series	Detects harm after exposure; testing and reporting are unequal across groups
Inspection and code-violation records	Home lead inspections; housing-code violations; fire inspections [8], [6], [12]	Lead, pests, dampness, fire, general housing quality	Building-level; records repeat findings at the same address	Reflects past enforcement choices (selection bias); inconsistent coding across jurisdictions
Resident complaints and digital traces	311 service requests; online reviews; search and location logs [20], [11], [15]	Pests, housing conditions, food hygiene	High volume; near real time; low cost	Propensity to report varies socio-spatially; privacy and platform-access constraints
Building and property registries	Construction year, assessor and ownership data, permits, cooling-tower registries [8], [14], [7]	Lead, *Legionella*, fire	Stable structural predictors; links units to owner portfolios	Incomplete or outdated records (e.g., service-line material)
Environmental measurements and sensors	Indoor temperature loggers; radon test databases; IAQ sensor networks [21], [23], [29]	Heat, radon, IAQ, dampness	Direct exposure measure; continuous monitoring possible	Sparse coverage; calibration drift; cost of deployment at scale

3.3 Risk Prediction and Inspection Targeting

Predictive targeting of inspections was the most advanced of the various methods. In Chicago, a model which had been trained on about 2.5 million blood-lead tests, 120,000 home-inspection records, building footprints, assessor data, and census characteristics was able to identify houses likely to have lead hazards before any child had been poisoned; the address-level features that proved most informative were building age, building condition, and the history of previous poisonings and inspections at that address, and the precision of the model was better than that obtained by random selection at all the levels of intervention examined [8]. A later validation that included an additional 2 million building permits and violations found a positive predictive value of 15.5% for detecting children at the highest risk—approximately twice the figure of earlier predictive methods—and showed comparable performance across different racial and ethnic groups [16]. Area-level random forest models, developed using socio-economic, housing, and water-quality data, achieved ROC AUCs of 0.91 in Massachusetts and 0.85 in New York [17], and the same method was used to identify schools at high risk of having elevated levels of lead in their drinking water from publicly available data alone [18], being one of the rare instances of such an approach being applied to institutional buildings. In Flint, Michigan, an adaptive statistical model was used to decide which homes should have their lead service lines excavated when the municipal records regarding pipe material were incomplete, and it updated its predictions as the results of the excavations came in [14].

In addition to lead, a fire-risk model used by the Atlanta Fire Rescue Department assigned risk scores to over 5,000 buildings, achieving true-positive rates as high as 71%, and detected 6,096 commercial properties that satisfied the inspection criteria but had not been included on the inspection list [12]. Studies of food premises demonstrate how digital traces can supplement administrative data: analysis of the text enabled predictions of hygiene violations [11], [13], and a model based on aggregated search and location data identified premises of which 52.3% were considered unsafe upon inspection, as compared with 24.7% in the case of baseline inspections [15]. Even though food premises are not the main focus of this review, these results show the increase in inspection yield that can be achieved through predictive targeting when it is verified against inspection outcomes.

3.4 Spatial and Cluster Analysis

The use of spatial methods showed where hazards tend to concentrate and allowed clusters to be identified which indicated a common and ongoing cause. At the level of the census tract, there was an association between the density of housing-code violations and the use of emergency departments and hospitals related to paediatric asthma [6], and in New York City the number of rat sightings reported through the 311 service was linked to the neighbourhood's sociodemographic characteristics, housing features, and distance from open space [20]. The most clear-cut instance of detecting a recurring hazard through clustering was a clinical rather than a statistical one: a medical-legal partnership noticed that 45 children with asthma, developmental delay, or elevated lead levels lived in 16 housing units owned by a single owner, and as a result of advocacy at the portfolio level systemic repairs were carried out in 19 building complexes [5]. Community heat-vulnerability indices, which are based on census data, land-cover data, and air-conditioning data, act as an area-level measure of heat [19]. Although prospective space-time scan statistics [31] and aberration-detection algorithms [32], which are widely used in infectious disease surveillance, were not applied to the building-hazard data in any of the studies included, they are directly appropriate for detecting emerging clusters of complaints, violations, or cases.

3.5 Measurement and Exposure Modelling

Studies based on measurement data estimate exposure directly, rather than deducing it from administrative proxies. Measurements of indoor temperature taken on a hourly basis in the homes of older people in Detroit found that outdoor temperature, solar radiation, and dew point together accounted for 38% of the variation in indoor temperature, and that indoor maximum temperatures were often higher than the outdoor ones [21]. With regard to radon, kernel regression mapping of a national measurement database [22] and an ensemble learning model which predicts monthly community-level concentrations [23] show how it is possible to extend sparse measurements to dwellings that have not been measured. Nowadays, low-cost sensor networks and Internet-of-Things monitoring systems make it possible to carry out continuous indoor measurements, but the quality of the data from these systems varies and field calibration is necessary [29], [30]. As for dampness and mold, there are well-established building-physics growth models which convert temperature and relative-humidity histories into an estimate of mold risk [33], [34]; linking these models to sensor data is a straightforward way of achieving early warning, but none of the studies included evaluated such a linkage in relation to health or inspection outcomes. Institutional environments such as schools, in which indoor pollutants and thermal conditions have been associated with occupant performance [35], are still under-represented in measurement studies.

3.6 Case- and Event-Driven Surveillance and Registries

For hazards which result in acute, reportable diseases, case surveillance continues to be the main method of detecting such cases, and its importance is based on being able to trace the cases back to a specific building source. A national framework for carbon monoxide poisoning combined data on mortality, visits to emergency departments, hospitalizations, and cases reported to the poison centre in order to estimate the burden [24]. Reviews of investigations into outbreaks of Legionnaires' disease in North America found deficiencies in the management of building water systems to be the common, preventable cause [25], and surveillance data from 20 states and a large metropolitan area demonstrated that health-care-associated cases occurred in many facilities, thus supporting the introduction of mandatory water-management programmed in hospitals and long-term care establishments [26]. The experience of New York City shows how event surveillance can be turned into a registry-based preventive system: following six community outbreaks between 2006 and 2015, which involved 213 cases and 18 deaths, the city made it necessary to register cooling towers, carry out inspections, carry out maintenance and perform regular *Legionella* testing; by May 2017, 3,909 of the registered towers (79%) had been inspected and 46 had *Legionella* levels above the action thresholds [7]. Such a registry functions as a denominator for high-risk building systems against which future recurrences can be monitored.

3.7 Data Quality, Bias, and Equity

Two studies have examined the data used by surveillance systems, and as the tendency of residents in different areas to make 311 complaints depends on the socio-economic and demographic characteristics of their neighborhoods, the approach of assigning inspections on the basis of complaints can systematically fail to detect hazards in some communities [27]. A method for estimating reporting delays in the absence of ground-truth data has shown that there are significant spatial and socio-economic differences in the speed at which incidents are reported in both New York City and Chicago [28]. The same issue arises with models based on blood-lead test data, which indicate who has been tested, and on inspection records, which show where inspectors have been sent in the past [16]. Experience in the health care sector demonstrates that algorithms trained on proxies influenced by unequal access can amplify that inequality on a large scale [36].

Table 3: Synthesis by method family: typical use, maturity of evidence, strengths, and principal gaps

Method family	Typical question answered	Evidence maturity (studies)	Strengths	Principal gaps
Risk prediction and inspection targeting	Which buildings or units should be inspected first?	Highest: 9 studies; 3 deployed	Measurable gains in inspection yield; uses existing administrative data	Concentrated on lead; few prospective health-outcome evaluations; label bias from unequal testing
Spatial and cluster analysis	Where do hazards concentrate, and do clusters share a cause?	Moderate: 4 studies; 1 deployed	Reveals owner- and neighbourhood-level recurrence; supports targeted enforcement	Mostly cross-sectional; prospective scan statistics not yet applied to building data
Measurement and exposure modelling	What is the actual exposure in this building?	Low: 3 studies; none deployed	Direct exposure measure; continuous monitoring feasible	Sparse coverage; sensor calibration; mold and IAQ models not linked to outcomes
Case- and event-driven surveillance, registries	Which building system caused these cases, and which systems are at risk?	Moderate: 4 studies; 1 deployed	Strong legal basis; registries define the population of high-risk systems	Detects harm after exposure; limited predictive use of registry data
Data-quality and equity audit	Who is missing from the data?	Emerging: 2 studies	Quantifies reporting bias that affects all other families	Rarely built into operational systems

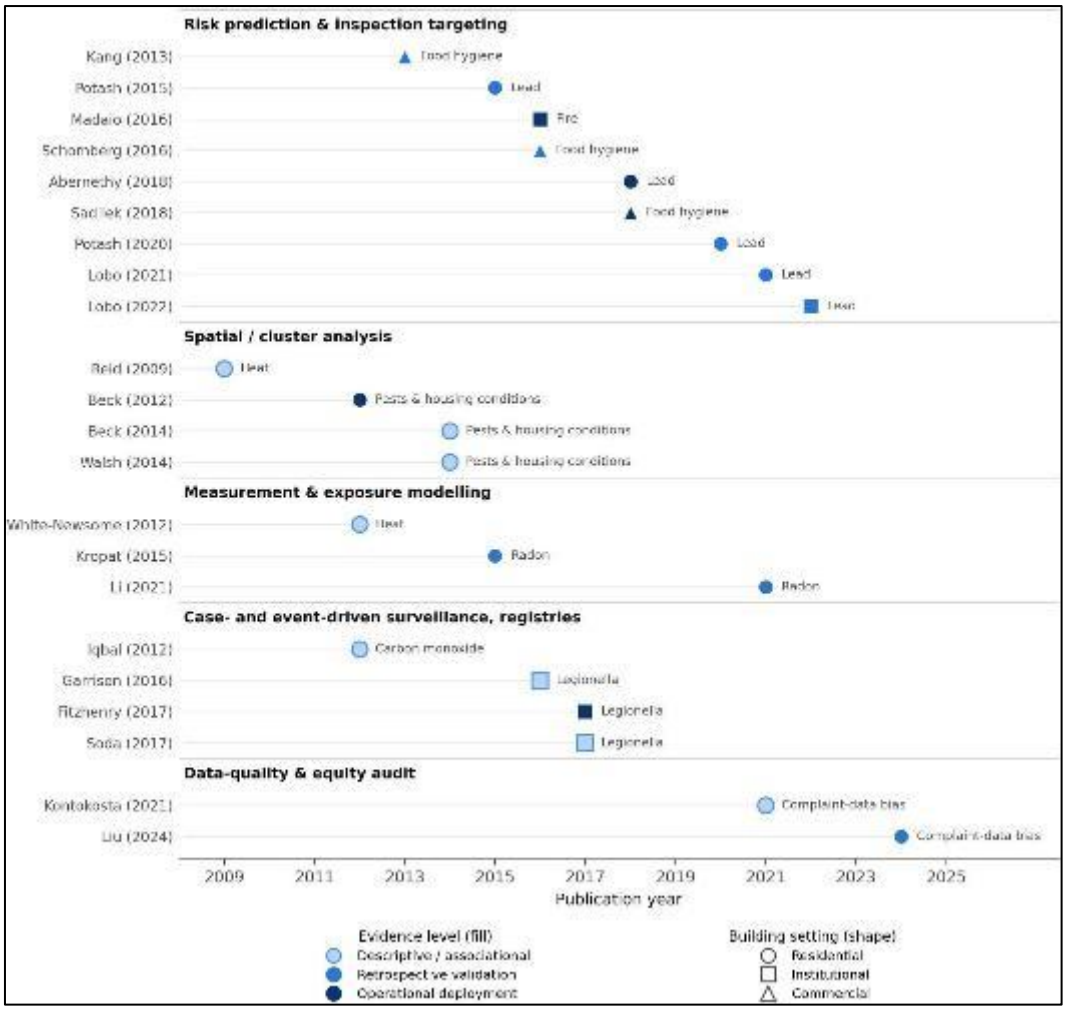


Figure 3. Included studies by publication year and method family. Fill indicates evidence level (descriptive, retrospective validation, or operational deployment); marker shape indicates building setting; the text beside each marker names the hazard domain. Operational deployments are few and scattered across families.

4 DISCUSSION

4.1 Principal Findings

This review found a small but growing body of evidence that data-driven methods can detect and anticipate recurring environmental hazards in buildings. The strongest evidence concerns predictive targeting of lead-hazard inspections, where models trained on surveillance, inspection, and building data have been validated on later data and shown to outperform existing approaches [8], [16], [17].

4.2 Detecting Recurrence

Improvements in inspection efficiency have also been observed for fire risk [12] and food-premise hygiene [15]. Registers of high-risk building systems, such as the regulation of cooling towers in New York City, transform repeated outbreaks into a form of preventive monitoring [7]. Yet, only five of the twenty-two studies described actual operational implementation. The evidence is limited to a small number of U.S. cities and relates to just one type of hazard, and most combinations of hazard and monitoring method have not been investigated (Figure 2).

4.3 Detecting Recurrence

Recurrence was almost never dealt with explicitly, even though it was the factor which made hazards predictable. The best predictors of lead hazards were the address-level records of previous poisonings and inspections.

The Cincinnati cluster only became apparent after cases were linked on the basis of ownership [5] and the cooling-tower registry functions since the same systems are tested repeatedly [7]. The findings suggest that, like a clinical registry, the building, the unit and the owner portfolio can be considered units of analysis in their own right in the context of a

building hazard surveillance system. Some of the practical needs are that the structure has a fixed identifier used by health, housing and environmental departments, that the address information is linked as well as the owner information, and that the outcomes are documented on repeated violations and repeated cases and not just the first.

4.4 Institutional Buildings

Institutional buildings were represented too little. The only studies that used predictive or sensor-based surveillance were those concerning school drinking water [18], cooling towers [7], and health-care-associated Legionnaires' disease [26]; no other included study applied such surveillance to schools, hospitals, or long-term care facilities. These buildings contain populations who are particularly vulnerable to the effects of heat, poor indoor air quality, and waterborne pathogens [21], [35], and many of them already have building-management systems and are required to have water-management plans which produce a continuous flow of data. There is a clear opportunity in linking those data streams to surveillance and anomaly-detection methods [37].

4.5 Equity and Governance

There is a selection mechanism for each data source in Table 2. The complaints that are recorded don't accurately represent all the neighborhoods [27] [28] because complaints are not received by all of them, testing data only cover what is being screened and inspection data only reflect where the inspectors are sent. Thus, the gaps that these data aim to fill can be reproduced using predictive models trained on them [36]. Fortunately, a study comparing the performance of lead prediction across different racial/ethnic groups reported similar results [16] and demonstrated the feasibility of lead-prediction audits. Routine subgroup evaluation, reporting bias corrected, a proportion of inspections carried out outside of the model scores to reduce feedback loops, and transparent governance, agreed to with affected communities should be included in deployment. These assessments can be easily made using the surveillance-evaluation characteristics of sensitivity, predictive value, representativeness and timeliness [10].

4.6 Strengths and Limitations

One of the strengths of this review is the way it brings together evidence from public health, building science, and computer science into a single framework (Figure 1) and the clear categorisation of the level of evidence. However, there are several limitations. First, the review was carried out by a single reviewer without the existence of a registered protocol, and the search process involved both database and web-index searches as well as citation chaining; it is therefore possible that relevant studies—especially those in the grey literature, municipal reports, and non-English publications—were overlooked, so the figures given here should be understood as referring to the sample of studies that were included rather than to the whole body of literature. Second, the studies included in the review used different outcomes and metrics, making it impossible to carry out a quantitative pooling. Third, the classification of the evidence level is based on the author's judgement and does not take the place of a formal risk-of-bias assessment. Fourth, publication bias is likely to favour cases of successful deployments. Fifth, the majority of the studies were carried out in the United States, which means that the findings may not be generally applicable to other regulatory and data environments.

4.7 Research Priorities

The priorities are as follows: studies that are prospective in nature and preferably randomised, with health outcomes used as the endpoint; the inclusion of explicit modelling of recurrence at both the building level and the owner-portfolio level; the use of prospective space-time cluster detection [31] for violation, complaint, and case data; the combination of low-cost sensor networks with building-physics models of mold and heat and the verification of these against inspection results [29], [34]; extending the approach to include schools, health care facilities and long-term care facilities; and the standard reporting of subgroup performance together with reporting-bias correction [28].

5 CONCLUSIONS

The physical nature of environmental hazards in homes and institutional buildings, along with their long-term presence and association with specific buildings and owners, makes them ideal candidates for data-driven monitoring. In the case of childhood lead poisoning, as illustrated in this evidence, the use of predictive models, spatial clustering, measurement-based exposure models and registers of high-risk building systems can increase the efficiency and timeliness of inspections and remedial work. In order to make this potential a reality in more settings, it will be important to create shared building identifiers and connect data from various agencies, to explicitly model the nature of hazard recurrence, to create the approach in institutional settings, and to conduct prospective evaluations that examine health outcomes and issues of equity.

REFERENCES

[1] Thacker, S. B., & Berkelman, R. L. (1988). Public health surveillance in the United States. *Epidemiologic Reviews, 10*, 164–190.

[2] McGeehin, M. A., Qualters, J. R., & Niskar, A. S. (2004). National environmental public health tracking program: Bridging the information gap. *Environmental Health Perspectives, 112*(14), 1409–1413.

[3] Krieger, J., & Higgins, D. L. (2002). Housing and health: Time again for public health action. *American Journal of Public Health, 92*(5), 758–768.

[4] Jacobs, D. E., et al. (2002). The prevalence of lead-based paint hazards in U.S. housing. *Environmental Health Perspectives, 110*(10), A599–A606.

[5] Beck, A. F., Klein, M. D., Schaffzin, J. K., Tallent, V., Gillam, M., & Kahn, R. S. (2012). Identifying and treating a substandard housing cluster using a medical-legal partnership. *Pediatrics, 130*(5), 831–838.

[6] Beck, A. F., Huang, B., Chundur, R., & Kahn, R. S. (2014). Housing code violation density associated with emergency department and hospital use by children with asthma. *Health Affairs, 33*(11), 1993–2002.

[7] Fitzhenry, R., et al. (2017). Legionnaires' disease outbreaks and cooling towers, New York City, New York, USA. *Emerging Infectious Diseases, 23*(11), 1769–1776.

[8] Potash, E., et al. (2015). Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)* (pp. 2039–2047). Association for Computing Machinery.

[9] Page, M. J., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71.

[10] German, R. R., et al. (2001). Updated guidelines for evaluating public health surveillance systems: Recommendations from the Guidelines Working Group. *MMWR Recommendations and Reports, 50*(RR-13), 1–35.

[11] Kang, J. S., Kuznetsova, P., Luca, M., & Choi, Y. (2013). Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1443–1448).

[12] Madaio, M., et al. (2016). Firebird: Predicting fire risk and prioritizing fire inspections in Atlanta. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 185–194). Association for Computing Machinery.

[13] Schomberg, J. P., Haimson, O. L., Hayes, G. R., & Anton-Culver, H. (2016). Supplementing public health inspection via social media. *PLOS ONE, 11*(3), e0152117.

[14] Abernethy, J., Chojnacki, A., Farahi, A., Schwartz, E., & Webb, J. (2018). ActiveRemediation: The search for lead pipes in Flint, Michigan. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '18)* (pp. 5–14). Association for Computing Machinery.

[15] Sadilek, A., et al. (2018). Machine-learned epidemiology: Real-time detection of foodborne illness at scale. *npj Digital Medicine, 1*, Article 36.

[16] Potash, E., et al. (2020). Validation of a machine learning model to predict childhood lead poisoning. *JAMA Network Open, 3*(9), e2012734.

[17] Lobo, G. P., Kalyan, B., & Gadgil, A. J. (2021). Predicting childhood lead exposure at an aggregated level using machine learning. *International Journal of Hygiene and Environmental Health, 238*, 113862.

[18] Lobo, G. P., Laraway, J., & Gadgil, A. J. (2022). Identifying schools at high-risk for elevated lead in drinking water using only publicly available data. *Science of the Total Environment, 803*, 150046.

[19] Reid, C. E., et al. (2009). Mapping community determinants of heat vulnerability. *Environmental Health Perspectives, 117*(11), 1730–1736.

[20] Walsh, M. G. (2014). Rat sightings in New York City are associated with neighborhood sociodemographics, housing characteristics, and proximity to open public space. *PeerJ, 2*, e533.

[21] White-Newsome, J. L., et al. (2012). Climate change and health: Indoor heat exposure in vulnerable populations. *Environmental Research, 112*, 20–27.

[22] Kropat, G., et al. (2015). Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland. *Science of the Total Environment, 505*, 137–148.

[23] Li, L., et al. (2021). Predicting monthly community-level domestic radon concentrations in the greater Boston area with an ensemble learning model. *Environmental Science & Technology, 55*(10), 7157–7166.

[24] Iqbal, S., Clower, J. H., King, M., Bell, J., & Yip, F. Y. (2012). National carbon monoxide poisoning surveillance framework and recent estimates. *Public Health Reports, 127*(5), 486–496.

[25] Garrison, L. E., et al. (2016). Vital signs: Deficiencies in environmental control identified in outbreaks of Legionnaires' disease—North America, 2000–2014. *MMWR Morbidity and Mortality Weekly Report, 65*(22), 576–584.

[26] Soda, E. A., et al. (2017). Vital signs: Health care-associated Legionnaires' disease surveillance data from 20 states and a large metropolitan area—United States, 2015. *MMWR Morbidity and Mortality Weekly Report, 66*(22), 584–589.

[27] Kontokosta, C. E., & Hong, B. (2021). Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustainable Cities and Society, 64*, 102503.

[28] Liu, Z., Bhandaram, U., & Garg, N. (2024). Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science, 4*(1), 57–65.

[29] Morawska, L., et al. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment International, 116*, 286–299.

[30] Saini, J., Dutta, M., & Marques, G. (2020). A comprehensive review on indoor air quality monitoring systems for enhanced public health. *Sustainable Environment Research, 30*, Article 6.

[31] Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine, 2*(3), e59.

[32] Farrington, C. P., Andrews, N. J., Beale, A. D., & Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society: Series A, 159*(3), 547–563.

[33] Hukka, A., & Viitanen, H. A. (1999). A mathematical model of mold growth on wooden material. *Wood Science and Technology, 33*(6), 475–485.

[34] Vereecken, E., & Roels, S. (2012). Review of mold prediction models and their influence on mold risk evaluation. *Building and Environment, 51*, 296–310.

[35] Mendell, M. J., & Heath, G. A. (2005). Do indoor pollutants and thermal conditions in schools influence student performance? A critical review of the literature. *Indoor Air, 15*(1), 27–52.

[36] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

[37] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), Article 15.