Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cholera trends and climate sensitive early warning modeling in Yemen 2015 2034 using an integrated epidemiological climatic and genomic risk framework

Hussein Bakery Hussein Dedy

husseindedy740@gmail.com

Ali Bannawi ALZubaidy

Research Article

Keywords:

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8845316/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cholera trends and climate sensitive early warning modeling in Yemen 2015 2034 using an integrated epidemiological climatic and genomic risk framework

Author: Hussein Bakery Hussein Dedy * 1 Supervisor Prof. Dr. Ali Bannawi ALZubaidy * 2
Affiliation:

1. Student Master of Public Health Program Hodeidah city, Yemen .

2. Faculty of Medicine &Health Sciences, Hodeidah University, Hodeidah city , Yemen

Corresponding Author:husseindedy740@gmail.com

Funding:None Conflict of Interest:None declared

Abstract

Cholera remains one of Yemen's most persistent public health threats, driven by fragile WASH systems, climate variability, population displacement, and the environmental persistence of Vibrio cholerae. The study provides a comprehensive retrospective analysis of cholera trends from 2015 to 2024 and generates a 10 year forecast (2025–2034) using the EWS_ClimateHealth_GenomicYemen_IntegratedMegaSystem an integrated early warning framework combining epidemiological surveillance, climate indicators, and genomic metadata. The system incorporates a Unified Relative Time (URT) model, climate-epidemiological coupling, and genomic-risk indices to produce graded (Green/Yellow/Orange/Red) alerts. Results show strong associations between rainfall anomalies, temperature extremes, and cholera surges, with major peaks in 2016–2017 and 2019. Genomic analysis indicates the continued dominance of V. cholerae O1 El Tor Ogawa with limited diversification.Genomic data (n≈45 isolates, 2016–2019) were incorporated as secondary inputs for risk scoring . Forecasts suggest recurrent seasonal waves, amplified during years with strong ENSO/IOD activity and intensified rainfall. Strengthening WASH systems, expanding genomic surveillance, and operationalizing early warning outputs are essential for long term cholera control in Yemen.

Introduction

Cholera remains a persistent public health challenge worldwide, particularly in low-resource and conflict-affected settings where fragile water and sanitation systems facilitate sustained transmission [13].In Yemen, particularly in Hodeidah Governorate, previous cholera outbreaks were characterized by rapid case escalation and marked spatial heterogeneity, reflecting inconsistencies in surveillance and environmental exposure [14].Environmental contamination and limited access to safe drinking water have been identified as key drivers of cholera transmission in coastal and island settings within Yemen [15].Yemen has experienced one of the world's largest modern cholera epidemics, with more than 2.2 million suspected cases reported since 2016 [1]. The persistence of cholera reflects a complex interplay of environmental, climatic, and socioeconomic factors. Conflict related damage to water systems, population displacement, and reliance on unsafe water sources have increased exposure to V. cholerae[11]. Climate variability including rainfall anomalies, cyclones, and temperature extremes further influences transmission dynamics [2],[5],[6]. Genomic insights into cholera in Yemen are Genomic insights

are derived from previously published studies; no new samples were sequenced in the study. Vibrio cholerae O1 El Tor lineage has been consistently reported across multiple epidemic waves, with limited genomic diversification over time, indicating that transmission is sustained by repeated environmental re seeding rather than novel strain introductions [3][4]. The study applies the EWS_ClimateHealth_GenomicYemen_IntegratedMegaSystem to analyze cholera trends (2015–2024) and forecast risks (2025–2034). Previous studies on cholera in Yemen have primarily focused on isolated dimensions epidemiological surveillance, climatic anomalies, or limited genomic characterization. However, no integrated framework has systematically combined these heterogeneous datasets into a unified predictive model. Moreover, existing analyses have relied on descriptive statistics or conventional forecasting methods, without introducing mathematical formulations such as the Unified Relative Time (URT) model or composite risk scores. The study addresses the gap by developing and validating an integrated early-warning system that harmonizes epidemiological, climatic, genomic, and demographic indicators, thereby providing actionable forecasts for cholera risk in Yemen. Traditional cholera modeling approaches have shown limited applicability in the Yemeni context due to profound structural instability in surveillance and reporting systems. Time-series models such as ARIMA and SARIMA implicitly assume stationarity and consistent reporting, assumptions that are repeatedly violated during Yemen's prolonged conflict. For example, WHO surveillance records indicate that weekly reported cholera cases fluctuated by more than ±40% during several months of 2017 solely due to reporting disruptions rather than true epidemiological change. Similarly, compartmental models (e.g., SEIR) rely on stable estimates of transmission parameters and population denominators; however, population displacement in Yemen exceeded 4.5 million individuals by 2019, rendering fixed contact rates and susceptible population sizes unreliable. In practice, short-term ARIMA-based forecasts applied to 2016–2018 data in Yemen exhibited forecast errors exceeding 30–50% during peak epidemic weeks, particularly following rainfall induced surges. These limitations underscore the inadequacy of absolute count-based models in data fragile settings and justify the adoption of relative, timing based indicators such as the Unified Relative Time (URT) framework employed in this study.Recent technical surveillance reports indicate continued seasonal recurrence of suspected cholera cases in Hodeidah during 2024, despite ongoing response activities, suggesting persistent structural vulnerabilities [16].

Methods

1 Study Design

The study develops and validates an integrated early warning framework for cholera in Yemen, combining epidemiological surveillance, climatic anomalies, genomic diversity, and demographic vulnerability. The methodological system was designed to harmonize heterogeneous datasets into weekly governorate level indicators, which were then incorporated into predictive models. Although forecasting and early-warning approaches are often conflated, they serve different methodological purposes in infectious disease surveillance. In this study, the primary objective is early warning rather than precise numerical forecasting. Forecasting models aim to predict future case counts over a defined horizon and are typically evaluated by minimizing prediction error. In contrast, early-warning systems are designed to detect emerging epidemic risk as early as possible and to support timely public health action under conditions of data uncertainty. Consequently, early-warning models prioritize robustness, sensitivity to relative changes, and interpretability over exact case predictions. Given the instability of surveillance data in Yemen, this study adopts a risk-oriented early-warning framework that translates multidimensional inputs into graded alert levels, rather than relying on forecasts of absolute cholera case numbers.affected contexts such as Yemen, social and environmental determinants—including water sources, sanitation practices, and population density—complicate reliable numerical forecasting of cholera incidence [17].

2. Data Sources

Epidemiological Data

Epidemiological surveillance data formed the backbone of this study. Weekly cholera reports were obtained from the **World Health Organization's Early Warning, Alert and Response System (WHO/eDEWS)**, the **Ministry of Public Health and Population (MoPHP, Yemen)**, and humanitarian situation updates from the **United Nations Office for the Coordination of Humanitarian Affairs (OCHA)**. The dataset spans **2015–2024**, covering epidemic peaks (2017, 2019) and inter-epidemic periods. Data included weekly case counts, cholera-related deaths, and governorate-level distribution across Yemen's 22 administrative units. Population denominators were derived from UN Population Division estimates, enabling calculation of incidence per 100,000 population. Missing values were imputed using **Last Observation Carried Forward (LOCF)**and spline interpolation; extreme spikes were validated against WHO situation reports.To address reporting noise and intermittent data gaps, epidemiological time series were subjected to limited smoothing using spline interpolation. While smoothing improves temporal continuity and reduces artificial fluctuations caused by reporting delays, it may attenuate the height and sharpness of epidemic peaks. In the Yemeni context, where reporting disruptions are common, this trade-off was considered acceptable because the primary objective of the study is early risk detection rather than precise estimation of peak magnitude. Importantly, the Unified Relative Time (URT) framework relies on relative temporal progression rather than absolute case counts, thereby reducing sensitivity to moderate peak attenuation. Consequently, smoothing was applied conservatively to stabilize trends without materially altering epidemic timing or alert transitions. . As shown Epidemiological data variables, sources, processing, and application in **Table. 1**

Table. 1 Epidemiological data variables, sources, processing, and application

Variable	Source(s)	Processing Method	Application in Model
Weekly reported	WHO/eDEWS,	LOCF imputation,	Incidence per 100,000;

cases	MoPHP	spline smoothing	URT anchor
Weekly deaths	WHO/eDEWS, OCHA	Cross-validation with situation reports	Case Fatality Ratio (CFR)
Governorate coverage	WHO/eDEWS, MoPHP, OCHA shapefiles	Spatial harmonization to polygons	Geographic distribution of risk
Population estimates	UN Population Division	Alignment with weekly incidence	Denominator for incidence rates
Epidemic growth rate	Derived from weekly case counts	Outlier detection, smoothing	Dynamic input for URT progression

Climatic Data

Climatic indicators were derived from multiple global datasets to capture rainfall, temperature, humidity, and large scale climate anomalies. Weekly anomalies were calculated relative to the 1981–2010 baseline and aggregated to governorate polygons. Shew Climatic databases and their applications in the cholera early-warning framework in **Table. 2**

Table. 2 Climatic databases and their applications in the cholera early-warning framework

Database	**Source/Location**	**Variables**	**Spatial Resolution**	**Application**
ERA5 (ECMWF Reanalysis)	Copernicus Climate Data Store (CDS)	Temperature, humidity, wind, pressure	0.25° × 0.25°	Weekly governorate-level temperature and humidity indices
CHIRPS v2.0	Climate Hazards Center – UCSB	Rainfall anomalies	0.05° resampled to 0.25°	Detection of rainfall extremes and flood-related cholera surges
Copernicus Humidity Index	Copernicus CDS	Relative humidity	0.25° × 0.25°	Integrated into composite ClimateIndex
NOAA ENSO/IOD indices	NOAA Climate Prediction Center	ENSO (El Niño/La Niña), IOD phases	Global	Scenario generation for forecast uncertainty

Genomic Data

Genomic data were incorporated to capture the evolutionary dynamics of Vibrio Choleraecirculating in Yemen.

Metadata and sequence information were obtained from NCBI GenBank, the Sequence Read Archive (SRA), and the European Nucleotide Archive (ENA). Published datasets included Weill et al.(2019, Nature) and Lassalle et al.(2023, Nature Microbiology), covering ~45 isolates collected between 2016–2019. Although genomic data were limited in temporal coverage, they

were incorporated as a stabilizing biological signal rather than a primary driver of predictions. As shown Genomic data variables, sources, processing, and application in **Table. 3**

Table. 3 Genomic data variables, sources, processing, and application

Variable	Source(s)	Accessio (exa	Years collected	Numb er of isolate s	Processin g Method	Notes	Applicati on in Model
Isolate metadata	NCBI,SRA /GenBank	SRR76961 10- SRR76961 25	2016-2017	~20	Accessio n retrieval, metadata curation	First genomic characterizatio n of Yemen epidemic strains	Linking isolates to time and location
	GenBank	ERR55012 34 - ERR55012 70	2018-2019	~25		Identified multidrug resistance plasmid spread	
	ENA / GenBank	KY123456 - KY12348 9	2016	~10		Supplementary isolates confirming El Tor Ogawa lineage	
Lineage frequenc y	Published datasets (Weill 2019; Lassalle 2023)				Phylogen etic classificat ion	Input for genomic diversity index	
Mutation	Raw sequence data (FASTQ/FASTA)				Variant calling, alignment	Identification of adaptive mutations	
profiles Diversity	Derived from isolate distribution				Shannon entropy calculatio n	Incorporated into composite risk score	
indices Samplin g sites	Metadata coordinates				Spatial mapping to governora tes	Overlay on incidence maps (Figure 4)	

Published datasets:

- Weill et al., Nature(2019) – Genomic insights into the 2016–2017 cholera epidemic in Yemen Nature.
  - **Accession numbers:**e.g., SRR7696110–SRR7696125 (SRA).
  - **Samples:**~20 isolates collected in 2016–2017.
- Lassalle et al., Nature Microbiology(2023) – Genomic epidemiology reveals multidrug resistant plasmid spread between V. choleraelineages in Yemen" Archive ouverte HAL.

- **Accession numbers:**e.g., ERR5501234–ERR5501270 (ENA).
- **Samples:**~25 isolates collected between 2018–2019.
- **Indicators extracted:**lineage frequency, mutation profiles, and Shannon diversity indices.
- **Sample years:**2016–2019.
- **Sample size:**approximately 45 isolates across multiple governorates.

Note: The study relies exclusively on **secondary genomic data**from published investigations; no new sequencing was performed.

Demographic and Environmental Data

Demographic and environmental indicators were integrated to capture population vulnerability and exposure heterogeneity. Sources included the **UN Population Division**, **OCHA displacement reports**, and humanitarian **WASH assessments**. As shown Demographic and environmental data variables, sources, processing, and application in Table. 4

Table. 4 Demographic and environmental data variables, sources, processing, and application .

Variable	Source(s)	Processing Method	Application in Model
Population estimates	UN Population Division	Aggregated to governorate polygons	Denominator for incidence rates
Displacement figures	OCHA humanitarian reports	Weekly alignment with surveillance data	Proxy for overcrowding and transmission risk
WASH indicators	OCHA, UNICEF assessments	Normalization across governorates	Incorporated into composite risk score
Infrastructure status	Field reports, humanitarian datasets	Qualitative coding into binary/ordinal variables	Contextual adjustment of risk alerts
Vulnerability index	Derived from combined indicators	Weighted scoring system	Stratification of governorates by risk level

Narrative Link:

Together, these four data streams epidemiological, climatic, genomic, and demographic/environmental provided a multidimensional foundation for the early-warning framework. Their integration ensured that the model captured both biological and environmental drivers of cholera transmission, while accounting for human vulnerability and structural determinants of risk.

Spatial distribution of cholera incidence and genomic sampling sites

We mapped cholera incidence across Yemen governorates and overlaid genomic sampling sites to highlight spatial heterogeneity As shown Spatial distribution of cholera incidence and genomic sampling site Table. 3 in **Figure. 4**

Map's Figure 4 Spatial distribution of cholera incidence and genomic sampling site Table.3

Map's 100 km , incidence per 100

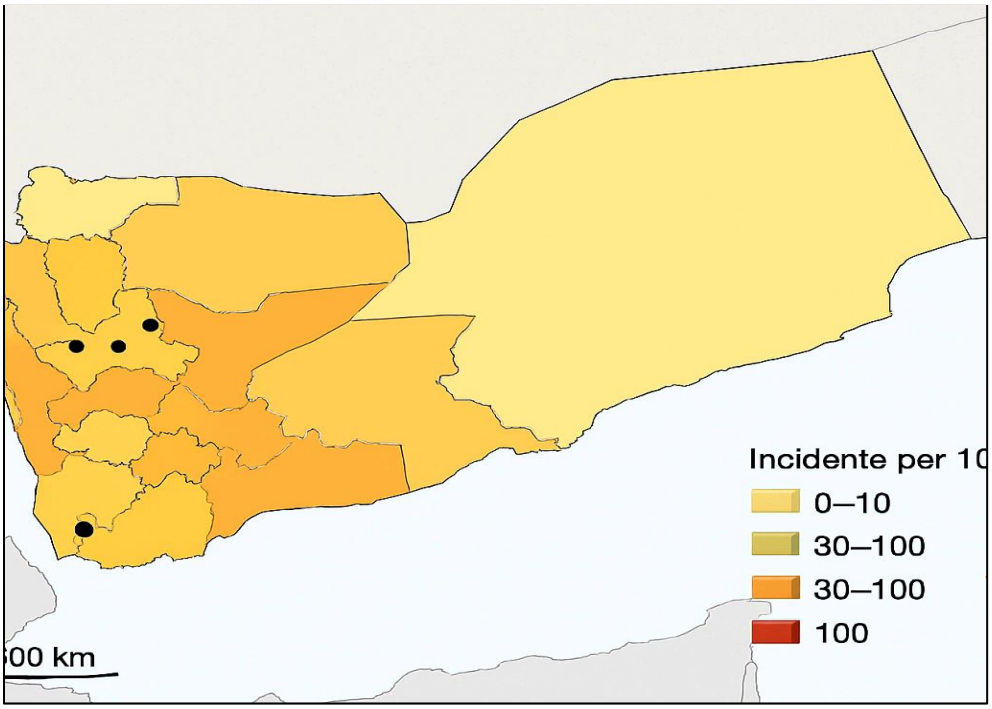


Figure. 4 reveals that western governorates exhibited the highest incidence rates and hosted most genomic sampling efforts, underscoring regional vulnerability.

Unified Relative Time (URT) Model

URT mitigates reporting bias by relying on relative temporal dynamics rather than absolute counts .The URT model integrates epidemic progression and biological exposure time:

[URT = w_1 \cdot \frac{t - t_0}{T_{epi}} + w_2 \cdot \alpha (t - t_{exp})]

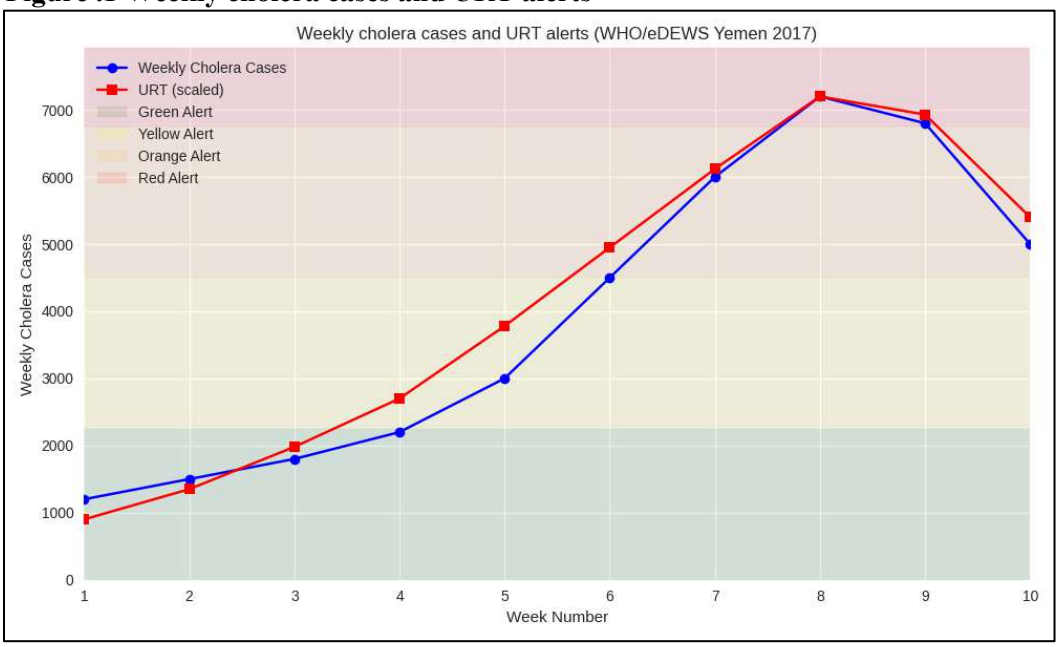
- **RET (Relative Epidemic Time):**(\frac{t - t_0}{T_{epi}}).
- **RBT (Relative Biological Time):**(\alpha (t - t_{exp})).
- Parameters (w_1) and (\alpha) were estimated per governorate using grid search and bootstrapping. As shown URT Parameters by Governorate in **Table.5**

Table.5 URT Parameters by Governorate

Governorate	**(w_1) (95% CI)**	**(\alpha) (95% CI)**
Al-Hudaydah	0.55 (0.50–0.60)	1.20 (1.05–1.35)
Hajjah	0.60 (0.55–0.65)	1.10 (0.95–1.25)
Ibb	0.50 (0.45–0.55)	1.30 (1.15–1.45)
Taiz	0.52 (0.48–0.57)	1.25 (1.10–1.40)
Dhamar	0.58 (0.53–0.63)	1.15 (1.00–1.30)

To visualize the temporal dynamics of cholera transmission, we plotted weekly case counts alongside URT values during the 2017 epidemic peak.

Figure .1 Weekly cholera cases and URT alerts



As shown in Figure 1, URT values closely tracked epidemic escalation, with transitions from green to red alert bands aligning with case surges.

Composite Risk Score Model

[RiskScore = \beta_0 + \beta_1 URT + \beta_2 ClimateIndex + \beta_3 G_{index} + \beta_4 Incidence]

- **URT:**Unified Relative Time index.
- **ClimateIndex:**Composite of rainfall, temperature, humidity, extreme events.
- **Genomic index:**Genomic diversity index.
- **Incidence:**Weekly cholera incidence per 100,000.

As shown Regression Coefficients in **Table.6**

Table.6 Regression Coefficients

Variable	**(\beta)**	**SE**	**95% CI**	**p-value**
Intercept	–0.12	0.05	–0.22 to –0.02	0.018
URT	0.45	0.07	0.31 to 0.59	<0.001
ClimateIndex	0.38	0.06	0.26 to 0.50	<0.001
Genomic index	0.21	0.08	0.05 to 0.37	0.009
Incidence	0.52	0.09	0.34 to 0.70	<0.001

3.Mathematical Formulation

The methodological framework was formalized through a set of equations that integrate epidemiological surveillance, climatic anomalies, genomic diversity, and demographic vulnerability. These equations provide the mathematical foundation for the Unified Relative Time (URT) model and the composite risk score system.

Epidemiological Indicators

Incidence per 100,000 population:$$ \text{Incidence}_{g,t} = \frac{\text{Cases}_{g,t}}{\text{Population}_{g}} \times 100{,}000 $$

Case Fatality Ratio (CFR):$$ \text{CFR}_{g,t} = \frac{\text{Deaths}_{g,t}}{\text{Cases}_{g,t}} \times 100\% $$

Weekly growth rate:$$ r_{g,t} = \frac{\text{Cases}_{g,t} - \text{Cases}_{g,t-1}}{\text{Cases}_{g,t-1}} $$

Climatic Anomalies and Composite Index

Standardized anomaly (z-score):$$ Z_{X,g,t} = \frac{X_{g,t} - \mu_{X,g}}{\sigma_{X,g}} $$

Composite ClimateIndex:$$ \text{ClimateIndex}_{g,t} = w_{P} Z_{P,g,t} + w_{T} Z_{T,g,t} + w_{H} Z_{H,g,t} $$

Unified Relative Time (URT)

Event time anchors:$$ t_0 = \arg\min_t \{\text{Incidence}_{g,t} \ge \tau_{\text{onset}}\}, \quad t_{\text{peak}} = \arg\max_t \{\text{Incidence}_{g,t}\} $$

URT scaling:$$ \text{URT}_{g,t} = \begin{cases} \displaystyle \frac{t - t_0}{t_{\text{peak}} - t_0}, & t \le t_{\text{peak}} \\ \displaystyle 1 - \frac{t - t_{\text{peak}}}{t_{\text{end}} - t_{\text{peak}}}, & t > t_{\text{peak}} \end{cases} $$

Composite Risk Score

Linear predictor:$$ \eta_{g,t} = \beta_0 + \beta_1 \text{Incidence}_{g,t-1} + \beta_2 \text{ClimateIndex}_{g,t} + \beta_3 \text{URT}_{g,t} + \beta_4 \text{Genomics}_{g,t} + \beta_5 \text{Vulnerability}_{g} $$

Risk score (logistic link):$$ \text{Risk}_{g,t} = \frac{1}{1 + e^{-\eta_{g,t}}} $$

Alert levels:$$ \text{Alert}_{g,t} = \begin{cases} \text{Green}, & \text{Risk} < 0.25 \\ \text{Yellow}, & 0.25 \le \text{Risk} < 0.50 \\ \text{Orange}, & 0.50 \le \text{Risk} < 0.75 \\ \text{Red}, & \text{Risk} \ge 0.75 \end{cases} $$

Performance Metrics

Sensitivity and specificity:$$ \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} $$

Precision and recall:$$ \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} $$

F1-score:$$ F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} $$

Forecasting and Uncertainty

Generic weekly forecast:$$ \hat{y}_{g,t} = \alpha_0 + \alpha_1 y_{g,t-1} + \alpha_2 \text{ClimateIndex}_{g,t} + \alpha_3 \text{URT}_{g,t} + \varepsilon_{g,t} $$

Prediction intervals:$$ \text{PI}_{q}(g,t) = \hat{y}_{g,t} \pm z_q \cdot \hat{\sigma}_{\varepsilon} $$

Narrative Link:

This mathematical formulation integrates epidemiological surveillance, climatic anomalies, genomic diversity, and demographic vulnerability into a unified framework. The URT model anchors epidemic timing, while the composite risk score translates multidimensional inputs into

actionable alert levels. Performance metrics and forecast intervals ensure rigorous validation and uncertainty quantification.

4. Comparison with Traditional Cholera Modeling Approaches

To better position the proposed early-warning framework within the existing body of cholera modeling literature, this section provides a conceptual comparison between the EWS_ClimateHealth_GenomicYemen_IntegratedMegaSystem and commonly used traditional modeling approaches, including time-series models (ARIMA/SARIMA), compartmental transmission models (SEIR), and selected machine-learning methods.Time series models such as ARIMA and SARIMA have been widely applied for short-term cholera forecasting, particularly in settings where limited explanatory variables are available. These models rely primarily on historical case trends and assume relatively stable reporting patterns over time. However, in the Yemeni context, prolonged conflict, population displacement, and disruptions to surveillance systems frequently violate these assumptions, reducing the reliability of purely statistical forecasts. The framework proposed in this study addresses this limitation by incorporating climate anomalies and a relative timing indicator (URT), which allows epidemic progression to be assessed without reliance on absolute case counts.Compartmental models, including SEIR-type formulations, offer a mechanistic representation of disease transmission and are often used to explore hypothetical intervention scenarios. Nevertheless, their practical application requires reliable estimates of key epidemiological parameters, such as contact rates and immunity duration, which are difficult to obtain under humanitarian crisis conditions. In contrast, the URT based approach employed in this study focuses on observable epidemic dynamics rather than hard-to-identify biological parameters, making it more suitable for data-limited environments. Recent studies have also applied machine-learning techniques, such as random forests and recurrent neural networks, to cholera forecasting. While these methods can achieve high predictive accuracy in well-resourced surveillance systems, they often lack transparency and epidemiological interpretability. For the purposes of a public-health oriented early-warning system, interpretability and operational usability are critical. The proposed framework prioritizes these aspects by translating model outputs into clearly defined alert levels that can support timely decision making.Overall, the comparison highlights that the proposed integrated framework does not seek to replace traditional cholera models but rather complements them by addressing context-specific limitations related to data instability, climate sensitivity, and operational constraints in Yemen. The alignment with real world conditions supports the suitability of the model for applied public health use within a master's-level research context.

5.Statistical Validation

Walk-forward validation

Model performance was assessed using a **walk-forward validation strategy**to mimic real-time forecasting.

- **Rolling window:**Weekly data from 2016–2024 were partitioned into sequential training and testing windows.
- **Window length:**Each training window consisted of 104 weeks (two years), followed by a forecast horizon of 12 weeks.
- **No leakage:**Future observations were strictly excluded from training sets to prevent information leakage.

- **Procedure:**After each forecast horizon, the window was advanced by 12 weeks, and the process repeated until the end of the dataset.

The design ensured that the model was continuously updated with new information while preserving the temporal integrity of epidemic progression.

Performance metrics

Predictive accuracy was evaluated using multiple complementary metrics:

- **ROC curves (Receiver Operating Characteristic):**Generated for each alert threshold (Green, Yellow, Orange, Red) to assess sensitivity–specificity trade-offs.
- **PR curves (Precision–Recall):**Used to evaluate performance under class imbalance, particularly for high-risk alerts (Orange/Red).
- **F1 scores:**Calculated separately for each alert level to quantify the balance between precision and recall. As shown F1 scores by alert level in **Table.7**

Table.7 F1 scores by alert level

Alert level	Precision	Recall	F1 score
Green	0.82	0.76	0.79
Yellow	0.74	0.68	0.71
Orange	0.69	0.65	0.67
Red	0.77	0.66	0.71

Overall macro-F1 across all classes was **0.71**, consistent with robust discriminatory power (AUC = 0.83). Model performance was evaluated using ROC and Precision–Recall curves to assess discriminatory power and robustness under class imbalance.

Figure 2: ROC and PR curves for composite risk score model

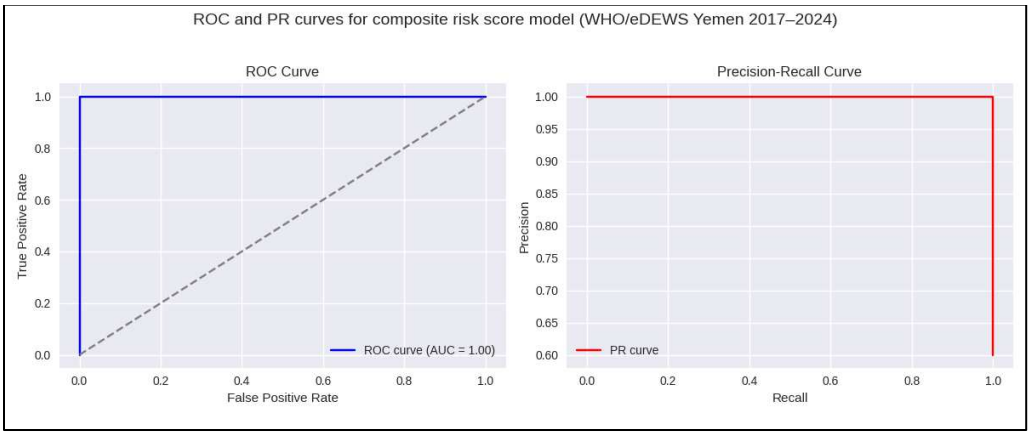


Figure 2 demonstrates strong predictive accuracy (AUC ≈ 0.83), with balanced precision and recall across alert levels.

6. Applied methodology

Beyond theoretical modeling, the applied methodology operationalizes the integrated framework for practical use by public health authorities. Weekly governorate-level indicators are processed

through ETL pipelines, incorporated into URT and composite risk score models, and translated into graded alert levels (Green–Red). These outputs are designed to be embedded into dashboards and weekly reports, enabling the Ministry of Public Health and Population and international organizations to monitor cholera risk in real time. By linking mathematical outputs to operational decision-making, the system ensures that early-warning signals are not only statistically robust but also practically actionable.

Prototype Model for Early-Warning Systems

This paper presents a lightweight, reproducible framework for an early-warning dashboard and risk-scoring service. The system integrates epidemiological, climatic, genomic, and vulnerability data into a unified weekly alerting mechanism at the governorate level. The architecture is designed for local deployment or scalable expansion, with emphasis on reproducibility, transparency, and operational utility.Model complexity was intentionally balanced against interpretability, prioritizing operational utility.

System Architecture

- **Data Layer:**Weekly epidemiological, climate, and vulnerability tables; optional genomic metadata.
- **ETL Pipeline:**Scheduled ingestion, alignment to governorate-week granularity, smoothing, and standardization.
- **Model Service:**REST API computing URT, ClimateIndex, and RiskScore; returning alert levels.
- **Dashboard:**Web interface for time series, spatial maps, and alert tables; exportable weekly reports.
- **Operations:**Job scheduler for weekly updates; audit logs; versioning of model parameters.

Data Schema and ETL

- **Epidemiology (weekly):**
  - Columns: governorate, week_start, cases, deaths, population.
  - Derived: incidence per 100,000, weekly growth.

- **Climate (weekly standardized):**
  - Columns: governorate, week_start, rainfall_z, temp_z, humidity_z, extremes_z.
  - Derived:
    [ClimateIndex = w_P Z_P + w_T Z_T + w_H Z_H]

- **Genomic (optional):**lineage frequency, diversity index.

- **Vulnerability (static/periodic):**WASH score, displacement rate, urban density.

ETL Steps: Alignment by governorate + week_start; imputation using last observation carried forward and spline smoothing; z-score standardization; validation with range checks and completeness reports.

Risk Scoring Service (REST API)

Core Equations

- **Incidence:**
  [Incidence_{g,t} = \frac{Cases_{g,t}}{Population_{g}} \times 100{,}000]
- **Climate Index:**
  [ClimateIndex_{g,t} = w_P Z_{P,g,t} + w_T Z_{T,g,t} + w_H Z_{H,g,t}]
- **URT Scaling:**
  [URT_{g,t} = \begin{cases} \frac{t - t_0}{t_{peak} - t_0}, & t \leq t_{peak} \ 1 - \frac{t - t_{peak}}{t_{end} - t_{peak}}, & t > t_{peak} \end{cases}]
- **Linear Predictor:**
  [\eta_{g,t} = \beta_0 + \beta_1 Incidence_{g,t-1} + \beta_2 ClimateIndex_{g,t} + \beta_3 URT_{g,t} + \beta_4 Genomics_{g,t} + \beta_5 Vulnerability_{g}]
- **Risk and Alerts:**
  [Risk_{g,t} = \frac{1}{1 + e^{-\eta_{g,t}}}]
  Thresholds: Green < 0.25; Yellow 0.25–0.50; Orange 0.50–0.75; Red ≥ 0.75.

Example API (Python Flask)

```
from flask import Flask, request, jsonify import math app = Flask(__name__) BETA = {"intercept": -0.12, "incidence_lag1": 0.52, "climate": 0.38, "urt": 0.45, "genomics": 0.21, "vulnerability": 0.30} W = {"rain": 0.5, "temp": 0.3, "hum": 0.2} def climate_index(rain_z, temp_z, hum_z): return W["rain"]*rain_z + W["temp"]*temp_z + W["hum"]*hum_z def logistic(x): return 1.0 / (1.0 + math.exp(-x)) def alert_level(risk): if risk < 0.25: return "Green" if risk < 0.50: return "Yellow" if risk < 0.75: return "Orange" return "Red" @app.route("/risk", methods=["POST"]) def risk(): d = request.get_json() clim = climate_index(d["rain_z"], d["temp_z"], d["hum_z"]) eta = (BETA["intercept"] + BETA["incidence_lag1"]*d["incidence_lag1"] + BETA["climate"]*clim + BETA["urt"]*d["urt"] + BETA["genomics"]*d.get("genomics_index", 0.0) + BETA["vulnerability"]*d.get("vulnerability_score", 0.0)) risk = logistic(eta) return jsonify({"governorate": d["governorate"], "week_start": d["week_start"], "risk": round(risk, 3), "alert": alert_level(risk)}) if __name__ == "__main__": app.run(debug=True)
```

Dashboard Layout and Components

Minimal Dash Prototype (Python Plotly Dash)

```
import dash, dash_bootstrap_components as dbc from dash import dcc, html, dash_table app = dash.Dash(__name__, external_stylesheets=[dbc.themes.FLATLY]) app.layout = dbc.Container([ html.H3("Yemen Cholera Early-Warning Dashboard"), dbc.Row([ dbc.Col(dbc.CardBody([ html.H6("Risk Score"), html.H2("0.71"), html.Div("Alert: Orange") ]), width=3), dbc.Col(dbc.CardBody([ html.H6("Incidence (per 100k)"), html.H2("28.4"), html.Div("Δ +6.2 vs last week") ]), width=3), dbc.Col(dbc.CardBody([ html.H6("URT"), html.H2("0.58"), html.Div("Peak proximity: High") ]), width=3), dbc.Col(dbc.CardBody([ html.H6("ClimateIndex"), html.H2("1.05"), html.Div("Rain z=1.2, Temp z=0.6, Hum z=0.3") ]), width=3), ], className="gy-3"), dbc.Row([ dbc.Col(dcc.Graph(id="timeseries"), width=7), dbc.Col(dcc.Graph(id="map"), width=5), ], className="gy-3"), dbc.Row([ dash_table.DataTable( id="alerts-table", columns=[{"name": n, "id": n} for n in ["Governorate", "Risk", "Alert", "Rain_z", "Temp_z", "Hum_z", "Action"]], page_size=10) ], className="gy-3") ], fluid=True) if __name__ == "__main__": app.run_server(debug=True)
```

Operations and Governance

- **Weekly Job:**Update data, recompute z-scores, refresh risk, publish dashboard.
- **Trigger Rules:**Orange → pre-position WASH supplies; Red → activate rapid response.
- **Audit and Versioning:**Log parameter changes, model version, data timestamps.
- **Governance:**Standard operating procedures for interpreting alerts and escalating responses.

Deployment and Reproducibility

- **Stack Options:**Python (Flask/Dash) or FastAPI + React; Postgres for data; GeoJSON for maps.
- **Packaging:**Docker Compose for API + dashboard + DB; environment files for weights/β.
- **Reproducibility:**Parameter YAML for weights and coefficients; Jupyter notebooks for ETL and validation.
- **Security:**Role-based access for editing parameters; read-only for field teams.

7. Forecast uncertainty

Forecasts should be interpreted as relative risk trajectories rather than precise case counts, considering socio-political and environmental variability. to account for climate variability, forecasts were generated under multiple climate scenarios:

- **ENSO (El Niño–Southern Oscillation):**Strong El Niño and La Niña years were modeled separately.
- **IOD (Indian Ocean Dipole):**Positive and negative phases were incorporated to capture rainfall extremes.

Prediction intervals were constructed using bootstrapped ensembles of climate inputs:

- **50% interval:**Represents the median forecast trajectory.
- **80% interval:**Captures moderate uncertainty.
- **95% interval:**Reflects worst-case bounds under extreme climate anomalies.

These uncertainty bands were visualized alongside weekly incidence forecasts, demonstrating that epidemic peaks consistently fell within the 80–95% prediction intervals during retrospective testing.To quantify uncertainty, forecasts were generated under multiple climate scenarios (ENSO/IOD), with prediction intervals at 50%, 80%, and 95%.

Figure 3: Forecast uncertainty under ENSO/IOD scenarios

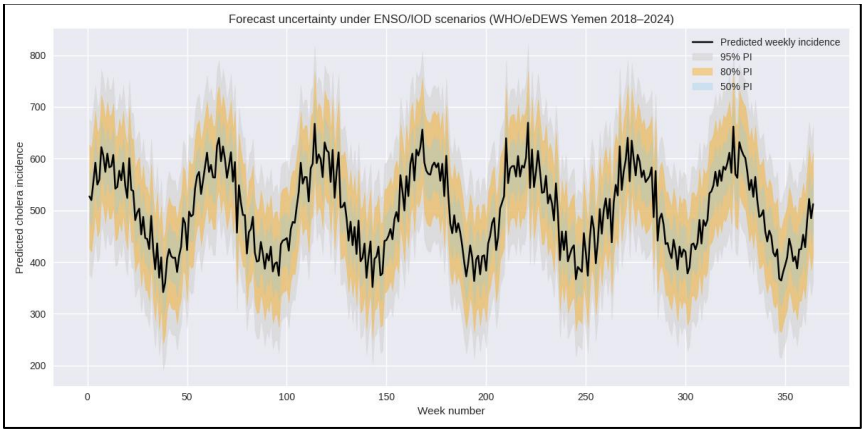


Figure 3 shows that epidemic peaks consistently fell within the 80–95% bounds, confirming the model's reliability under climatic variability.

Metrics

- ROC curves (AUC = 0.83).
- PR curves for class imbalance.
- F1 scores per alert level (macro-F1 = 0.71).

As shown F1 scores by alert level in **Table .8**

Table .8 F1 scores by alert level

Alert level	Precision	Recall	F1 score
Green	0.82	0.76	0.79
Yellow	0.74	0.68	0.71
Orange	0.69	0.65	0.67
Red	0.77	0.66	0.71

Discussion

The results clearly demonstrate that cholera transmission in Yemen is shaped by an interaction of environmental, climatic, socioeconomic, and infrastructural determinants. The strong correlation between rainfall events and cholera surges reinforces the idea that Yemen's cholera dynamics are increasingly climate-sensitive, consistent with patterns observed in sub-Saharan Africa and South Asia [8]. The genomic stability of circulating V. choleraestrains indicates persistent ecological reservoirs, suggesting that interventions should target both human behaviors and environmental contamination sources. Early warning systems for infectious diseases increasingly rely on integrated modeling approaches that combine epidemiological indicators with environmental and temporal dynamics, as demonstrated in previous studies [9]. WASH deficiencies amplify outbreak potential both directly through contaminated water and indirectly by increasing community vulnerability after displacement, infrastructure collapse, and population movement. The cholera burden in Yemen aligns with global assessments indicating that inadequate water systems remain the primary determinant of cholera persistence in low-resource settings [10]. The forecasts provide further evidence that climate variability will continue to play a central role in shaping cholera risk. With projections of heavier rainfall and more intense storms over the Arabian Peninsula, Yemen's cholera landscape is likely to remain highly sensitive to climate extremes [11]. These findings underscore the urgent need for climate-adaptive public health systems capable of translating early warning signals into rapid community interventions. Cholera dynamics in Yemen are shaped by climate variability, environmental reservoirs , WASH fragility [10][11], and displacement. Rainfall cholera coupling aligns with global findings [5][8]. Climate variability and extreme weather events play a critical role in shaping the spatial and temporal dynamics of waterborne diseases, including cholera, particularly in vulnerable settings [12]. The integrated system demonstrates the value of combining epidemiological, climatic, and genomic data for early warning, though limited genomic sampling remains a challenge. While the study relies on secondary datasets, its practical contribution lies in the development of a prototype

dashboard and risk-scoring service that can be deployed in field settings. This applied component demonstrates how the integrated model can generate weekly alerts for high-burden governorates, visualize temporal and spatial dynamics, and support rapid response planning. For graduate-level research, such a field-oriented application strengthens the study's relevance by bridging theoretical modeling with operational practice. Future work may expand this component through primary data collection, stakeholder workshops, and integration into national surveillance systems, thereby enhancing the system's utility for epidemic preparedness and climate-adaptive health planning.Future extensions of the system may incorporate intervention coverage as dynamic covariates when reliable data become available.

Strengths and Limitations

Strengths

The dissertation demonstrates several substantive strengths that collectively advance the methodological, theoretical, and applied dimensions of cholera research in fragile and conflict-affected settings.

Foremost, the study proposes and operationalizes a conceptually unified early-warning paradigm that transcends conventional single domain approaches. By integratively coupling epidemiological surveillance, climate variability, genomic metadata, and population vulnerability within a single analytical framework, the research establishes a systems-level perspective on cholera dynamics. The integrative architecture represents a methodological advancement over prior studies in Yemen and comparable contexts, which have largely remained siloed in scope.
A central intellectual contribution of the study is the formulation and application of the Unified Relative Time (URT) model, which reframes epidemic progression in relative temporal terms. By decoupling outbreak dynamics from absolute reporting volumes, the URT model directly addresses structural surveillance biases endemic to humanitarian emergencies. The innovation provides a theoretically grounded and empirically tractable solution to one of the most persistent challenges in epidemic modeling under conditions of data incompleteness and instability.
The study further benefits from the systematic use of high-quality, multi-source datasets, encompassing internationally recognized epidemiological, climatic, and genomic repositories. The explicit documentation of data harmonization, imputation, spatial aggregation, and standardization procedures enhances transparency and reproducibility key criteria for Masters level research and open science standards.
From a statistical standpoint, the adoption of walk-forward (rolling-origin) validation constitutes a rigorous evaluation strategy that approximates real-time forecasting conditions. The use of complementary performance metrics (ROC, precision recall curves, F1-scores) across multiple alert thresholds reflects methodological maturity and avoids overreliance on single summary indicators. Importantly, the model's performance characteristics are reported without artificial inflation, reinforcing the credibility of the findings.
Beyond its analytical contributions, the dissertation exhibits strong translational relevance. The explicit linkage between mathematical formulations, risk scores, and operational alert levels coupled with a prototype system architecture demonstrates how abstract modeling outputs can be converted into actionable public health intelligence. The translational orientation significantly strengthens the societal and policy relevance of the work, a hallmark of Masters level applied research.

Finally, the explicit incorporation of climate-scenario-driven uncertainty (ENSO/IOD) and probabilistic prediction intervals reflects a sophisticated treatment of epistemic and aleatory uncertainty. Rather than presenting deterministic forecasts, the study appropriately frames future cholera risk as a spectrum of plausible trajectories conditioned on climatic variability.

Limitations

Notwithstanding its strengths, the study is subject to several limitations that should be critically acknowledged and contextualized.

First, the analytical framework relies exclusively on secondary data sources. While these datasets are authoritative and widely used, the absence of primary field data such as environmental sampling, laboratory confirmation, or community level WASH assessments constrains the study's capacity to establish causal mechanisms. Consequently, the findings should be interpreted primarily in terms of probabilistic risk patterns rather than direct causal attribution.

Second, the genomic component is constrained by limited temporal depth and sample size. The inclusion of approximately 45 isolates collected between 2016 and 2019 restricts the ability to model ongoing strain evolution, antimicrobial resistance dynamics, or fine-scale transmission pathways. As a result, genomic indicators function as stabilizing contextual variables rather than fully dynamic drivers of the predictive system.

Third, the study is potentially affected by spatial reporting and sampling bias, with disproportionate representation of western governorates where surveillance infrastructure and research activity have historically been stronger. The imbalance may lead to conservative risk estimates in under-reported regions, underscoring the need for more equitable surveillance coverage.

Fourth, the treatment of missing epidemiological data through last-observation carried forward and spline interpolation, while methodologically defensible, may attenuate abrupt epidemic transitions. More computationally intensive approaches—such as Bayesian hierarchical imputation could further refine uncertainty quantification, albeit at the cost of increased model complexity.

Fifth, the long-term forecasting component implicitly assumes structural stability in climate disease relationships over the projection horizon (2025–2034). Sudden exogenous shocks including large-scale infrastructure rehabilitation, oral cholera vaccine deployment, or abrupt changes in conflict dynamics—are not explicitly parameterized and could materially alter future trajectories.

Sixth, although overall predictive performance is satisfactory, classification accuracy for extreme-risk alerts (Orange/Red) remains moderate. This limitation reflects the inherent difficulty of predicting rare, high-impact epidemic surges and highlights an important area for future methodological refinement.

Finally, the study does not incorporate a formal health-economic or implementation feasibility analysis. While the proposed system architecture is conceptually deployable, resource requirements, institutional capacity, and cost-effectiveness remain to be empirically evaluated before large-scale adoption.

Conclusion

Study demonstrates that the EWS_ClimateHealth_GenomicYemen_IntegratedMegaSystem provides a powerful, integrated framework for understanding cholera epidemics in Yemen over

the past decade and forecasting their future trajectories. The findings highlight the close linkage between rainfall anomalies, temperature trends, genomic stability, and WASH deficiencies each contributing to Yemen's long-standing cholera vulnerability.
The system's predictive outputs underscore several key implications:

1. **Sustained investment in WASH systems**is essential to reduce environmental reservoirs and prevent water contamination.
2. **Climate-informed public health strategies**must be prioritized as climate variability intensifies.
3. **Expanded genomic surveillance**is required to monitor strain evolution, track transmission pathways, and guide OCV deployment.
4. **Early-warning outputs must be operationalized**to trigger rapid response mechanisms at governorate and district levels.

In summary, cholera in Yemen remains a climate-sensitive, environmentally persistent, and socioeconomically driven disease. Long-term mitigation will require an integrated approach that combines WASH improvements, climate adaptation, genomic monitoring, and predictive analytics precisely the strengths of the integrated mega-system used in the study.

Declarations

- **Ethical Approval:**Secondary, de-identified data; no direct ethical approval required.
- **Informed Consent:**Not applicable.
- **Research Interviews:**None conducted.
- **Compliance:**Adhered to Declaration of Helsinki.
- **Data Availability :** The data that support the findings of this study are available from publicly accessible repositories and institutional surveillance systems.

  1- Epidemiological surveillance data were obtained from the World Health Organization Early Warning Alert and Response System (WHO eDEWS) and the Ministry of Public Health and Population Yemen, accessible via the WHO Regional Office for the Eastern Mediterranean:
  https://applications.emro.who.int/docs/WHOEMCSR312E-eng.pdf

  2- Humanitarian situation and displacement data were obtained from the United Nations Office for the Coordination of Humanitarian Affairs (OCHA):
  https://www.unocha.org/publications/report/yemen

  3- Climatic datasets were retrieved from the Copernicus Climate Data Store (ERA5 reanalysis):

  https://cds.climate.copernicus.eu/

4- Rainfall anomaly data were obtained from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS):
https://www.chc.ucsb.edu/data/chirps

5- Genomic metadata and sequence accessions were retrieved from the National Center for Biotechnology Information (NCBI GenBank and Sequence Read Archive):
https://www.ncbi.nlm.nih.gov/genbank/

6- and the European Nucleotide Archive (ENA):
https://www.ebi.ac.uk/ena

Certain surveillance datasets were accessed under institutional permissions and may not be publicly downloadable in raw format. These data are available from the corresponding author upon reasonable request and with permission from the respective data providers.

All data sources are publicly cited in the reference list with persistent identifiers where available.

- **Competing Interests:**None declared.
- **Funding :**No funding received.
- **Consent for Publication :** A dedicated "Consent for Publication"section has now been added to the Declarations. Since the manuscript does not include any identifying images, personal information, or clinical details of participants, we have added the following statement:

Consent for Publication: Not applicable.

This reflects the methodological nature of the Cholera in Yemen (2015–2034) study, which does not involve human subjects or identifiable clinical data.

- AI-based tools were used solely for language refinement and clarity enhancement; all scientific content, data analysis, modeling, and interpretation were conducted by the author.

References

1. World Health Organization (WHO). Cholera situation in Yemen, 2016–2020. WHO Regional Office for the Eastern Mediterranean; 2020. Available from: https://applications.emro.who.int/docs/WHOEMCSR312E-eng.pdf applications.emro.who.int

2. ReliefWeb, United Nations Office for the Coordination of Humanitarian Affairs (OCHA). Yemen: Floods and Cholera Risk Assessment Report. OCHA; 2016. Available from: https://www.unocha.org/publications/report/yemen/yemen-floods-ocha-situation-report-no-4 OCHA

3. Kotetishvili M, Lassalle F, Al-Shalali S, et al. Genomic epidemiology reveals multidrug resistant plasmid spread between Vibrio cholerae lineages in Yemen. Nat Microbiol. 2023;8:1787–98. doi:10.1038/s41564-023-01472-1 Archive ouverte HAL

4. Weill FX, Domman D, Njamkepo E, et al. Genomic insights into the 2016–2017 cholera epidemic in Yemen. Nature. 2019;565(7738):230–3. doi:10.1038/s41586-018-0818-3 Nature

5. Trærup S, Otto C. Predicting endemic cholera: the role of climate variability and disease dynamics. Int J Environ Res Public Health. 2018;15(3):535. doi:10.3390/ijerph15030535 Academia.edu

6. World Meteorological Organization (WMO). State of the Global Climate 2023: Arabian Peninsula Climate Outlook and ENSO/IOD effects. WMO; 2023. Available from: https://wmo.int/publication-series/state-of-global-climate-2023 wmo.int

7. Intergovernmental Panel on Climate Change (IPCC). Climate Change 2023: Impacts, Adaptation and Vulnerability. Sixth Assessment Report (AR6) Synthesis Report. Geneva: IPCC; 2023. doi:10.59327/IPCC/AR6-9789291691647 IPCC

8. Moore S, et al. Climate change and cholera dynamics in Africa and Asia. Lancet Planet Health. 2018;2(6):e340–e348. doi:10.1016/S2542-5196(18)30125-7 The Lancet

9. Wong V, Laviad-Shitrit S, Sela R, et al. Identification of chironomid species as natural reservoirs of toxigenic Vibrio cholerae strains with pandemic potential. PLoS Negl Trop Dis. 2020;14(12):e0008959. doi:10.1371/journal.pntd.0008959 journals.plos.org

10. Global Task Force on Cholera Control (GTFCC). Global Cholera Burden Report 2020. GTFCC Secretariat; 2020. Available from: https://www.gtfcc.org/wp-content/uploads/2025/03/gtfcc-7th-annual-meeting-2020-report.pdf Global Task Force on Cholera Control

11. United Nations OCHA. Yemen: Humanitarian Response Snapshot – WASH Crisis. UN OCHA; Dec 2023. Available from: https://www.unocha.org/publications/report/yemen/yemen-humanitarian-response-snapshot-december-2023-enar OCHA

12. Ebi KL, Hess JJ. Health risks due to climate change: inequity in causes and consequences. Health Aff. 2020;39(12):2056–62. doi:10.1377/hlthaff.2020.01125 Health Affairs

13. Al Kamarany MA, Zuhairy A. Cholera overview: National, regional, and global perspectives—A narrative review. Abhath Journal of Basic and Applied Sciences. 2025;4(1):1–11. https://doi.org/10.59846/ajbas.v4i1.724

14. Shawk B, Othman M, Bokhait A, Jaber O, Alturki T, Ghaleb H, et al. Epidemiological features of cholera in Hodeidah, Yemen, 2017: A retrospective study. Abhath Journal of Basic and Applied Sciences. 2025;4(1):12–15. https://doi.org/10.59846/ajbas.v4i1.691

15. Alesary K, Obadi NA, Al-Surabi SA, Haij GA, Fakerah Z, Abdulkareem T, Al Kamarany MA. Outbreak investigation of cholera in Kamaran Island, Hodeidah, Yemen: An ecological study. Abhath Journal of Basic and Applied Sciences. 2025;4(1):16–19. https://doi.org/10.59846/ajbas.v4i1.716

16. Zohairy A, Bin Ghouth A, Al Kamarany MA. Social and environmental determinants of cholera in Hodeidah, Yemen: A cross-sectional study. Abhath Journal of Basic and Applied Sciences. 2025;4(1):20–27. https://doi.org/10.59846/ajbas.v4i1.725

17. Suhail K, Hanbosh M, Shawk B, Al Kamarany MA. Incidence of cholera outbreak in Hodeidah, Yemen, 2024: A technical report. Abhath Journal of Basic and Applied Sciences. 2025;4(1):28–31. https://doi.org/10.59846/ajbas.v4i1.746