# An Agentic LLM-Powered Framework for Cholera Risk Prediction and Actionable Insights from Explainable ML and Statistical Insights

# **Abstract**

Interpreting the complex and multifactorial risk factors driving Cholera outbreaks remains a critical challenge for public health, particularly across diverse environmental and socio-economic contexts. This paper presents an integrated agentic framework that combines explainable machine learning (ML), statistical analysis, and a language model-powered question-answering system to support Cholera risk interpretation and public health decision-making. Using a multi-country dataset spanning 2000–2025, the framework applies three interpretable ML models—Explainable Boosting Machines (EBM), Natural Gradient Boosting (NGBoost), and TabNet—to predict Cholera incidence based on environmental, socio-economic, and infrastructural variables. In parallel, statistical methods including Pearson and Spearman correlation, and multivariate linear regression are used to validate and quantify associations between predictors and disease outcomes.

A LangChain-powered agent, implemented with LangGraph, is integrated into the system to interpret model outputs, analyze tabular results, and generate expert-like responses to natural language queries. The agent draws evidence from multiple CSV-based analyses—including feature importance scores, correlation matrices, regression coefficients, and model performance comparisons—to provide grounded, interpretable answers and policy recommendations. A Streamlit interface enables interactive exploration of Cholera risk factors by researchers, health professionals, and policy stakeholders.

Results show strong agreement among models on key predictors, such as rainfall frequency, stagnant water presence, and open defecation, with statistically significant relationships confirmed through regression analysis. The EBM model achieved the lowest RMSE (0.421), indicating superior predictive performance. This work demonstrates how explainable AI and LLM agents can be combined into a transparent, interpretable, and actionable framework for public health analytics, offering valuable insights for data-driven disease prevention strategies.

# **1.0 Introduction**

Cholera remains a significant public health challenge, especially in low-resource settings where water sanitation and healthcare infrastructure are limited. The incidence of Cholera is influenced by a complex interplay of environmental, socio-economic, and infrastructural factors, making it difficult to fully understand, predict, and mitigate outbreaks [1, 2]. Traditional epidemiological analyses and black-box predictive models often struggle to capture these multifaceted relationships transparently, limiting their utility for policymakers and health practitioners.

Recent advances in interpretable machine learning (ML) provide promising methods to uncover key risk factors with explainability, enabling more trustworthy insights. However, the diverse outputs from multiple ML models and statistical techniques can be difficult to integrate and communicate effectively. Leveraging agentic reasoning powered by large language models (LLMs) offers an innovative way to bridge this gap by facilitating natural language-based exploration and explanation of complex data.

## **1.1 Research Questions**

To guide this study, we investigate the following core research questions:

* How can an LLM-powered agentic framework integrate explainable machine learning and statistical reasoning to identify key risk factors for Cholera incidence across multiple regions?

This question underpins our effort to design a transparent, interactive, and actionable framework that combines predictive modelling, statistical validation, and natural language reasoning.

## **1.2 Contributions**

This paper makes the following contributions:

* Development of a multi-model analytical pipeline leveraging Explainable Boosting Machines (EBM), Natural Gradient Boosting (NGBoost), and TabNet to analyze Cholera risk factors.
* Integration of a LangChain-powered LLM agent that interprets model outputs and statistical analyses, enabling natural language question-answering for decision support.

# **2.0 Related Work**

The integration of machine learning (ML) in public health, particularly for infectious disease modelling, has grown considerably in recent years. However, the application of interpretable models combined with large language models (LLM)-based agents remains limited in the context of Cholera and similar waterborne diseases. This section reviews the current landscape across three core dimensions: machine learning in epidemiology, explainable machine learning (XAI) in healthcare, and the emerging use of LLM agents for question-answering and knowledge synthesis.

## **2.1 Machine Learning for Epidemiological Modelling**

ML methods have been increasingly applied to model and predict disease outbreaks by learning patterns from high-dimensional data. Supervised learning algorithms such as Random Forest, XGBoost, and Support Vector Machines have been used to predict Dengue, Malaria, and COVID-19 incidences with notable accuracy (Adnan et al., 2021; Xu et al., 2020). These models can capture complex non-linear relationships between predictors and outcomes, making them suitable for diseases influenced by a range of environmental and socio-economic factors. In Cholera research specifically, statistical models such as logistic regression and time-series models have historically been used to forecast outbreaks (Mukandavire et al., 2011). However, recent studies have begun to explore the use of ensemble-based ML models for spatial and temporal prediction of Cholera risk (Bain et al., 2022). Despite these advances, most of these efforts focus solely on predictive accuracy rather than interpretability or explainability.

## **2.2 Explainable Machine Learning in Health Contexts**

The field of explainable AI (XAI) addresses the "black-box" nature of many ML models, offering tools to understand and trust model outputs. Techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and interpretable models like Explainable Boosting Machines (EBMs) have gained traction in healthcare for understanding risk factors and treatment decisions (Caruana et al., 2015; Lundberg et al., 2017).

In disease epidemiology, explainable models have helped reveal the relative influence of social determinants, environmental exposures, and behavioural practices on disease outcomes (Holzinger et al., 2019). EBMs, in particular, are well-suited for policy-relevant domains as they provide intelligible feature-by-feature risk curves and global feature importance metrics. However, few studies incorporate multiple explainable models (e.g., EBM, NGBoost, TabNet) in comparative frameworks, and even fewer link those insights to dynamic policy generation.

## **2.3 LLM Agents for Knowledge Synthesis and QA**

Large Language Models such as GPT-4 and PaLM have demonstrated remarkable abilities in synthesizing information, answering domain-specific questions, and summarizing complex data. Frameworks like LangChain and LangGraph have made it possible to build agents that integrate LLM reasoning with structured data sources, allowing AI to perform contextual Q&A over documents, spreadsheets, and databases (Lewis et al., 2023). In public health, LLM agents are beginning to be tested for real-time decision support, clinical triage, and research summarization (Singhal et al., 2022). However, the use of LLMs to interpret and communicate the results of statistical and ML modelling—particularly in epidemiological settings—remains underexplored.

While ML and XAI methods have advanced significantly, and LLMs offer promising capabilities for interaction and synthesis, there is a notable **lack of integrated systems** that:

* Combine interpretable ML modelling with statistical reasoning,
* Provide real-time, data-driven explanations via a conversational interface,
* Translate modelling results into actionable, localized public health recommendations.

This study addresses this gap by proposing an **agentic framework** that unifies explainable ML models (EBM, NGBoost, TabNet), statistical correlation analysis, and an LLM-powered interface for Cholera risk analysis and policy guidance.

# **3.0 Dataset Description**

The dataset used in this study was obtained from Kaggle and is titled *Water Pollution and Disease*. It is a recently published dataset, uploaded in March 2025, and covers the period from 2000 to 2025, offering a contemporary and comprehensive view of factors influencing waterborne diseases, particularly Cholera. The dataset includes a wide range of variables spanning environmental indicators (such as rainfall frequency, water quality metrics), socio-economic factors (including sanitation practices, education levels), and health outcomes, specifically *Cholera Cases per 100,000 people*. The dataset is publicly available at [Kaggle Water Pollution and Disease Dataset](https://www.kaggle.com/datasets/khushikyad001/water-pollution-and-disease/data).

## **3.1 Variables**

* **Target Variable:** *Cholera Cases per 100,000 people* — the primary outcome measure used for modelling.
* **Predictor Variables:** 
  + Environmental: rainfall frequency, water source type, contaminant levels, pH, turbidity, dissolved oxygen, nitrate and lead concentrations, bacterial counts.
  + Socio-economic: sanitation access, open defecation rates, education levels, population density, and related infrastructural variables.

## **3.2 Data Preprocessing**

Initial validation confirmed the presence of the target variable in the dataset. Features were categorized into numerical and categorical types, excluding the target from transformation steps. To address skewness in feature distributions, the skewness coefficient was computed, and variables with a skew greater than 1 were log-transformed using the log1p function. This reduced the impact of outliers and improved the stability of downstream models. Numerical features were then normalized using **z-score standardization** to ensure consistent scale across variables. For categorical variables, **one-hot encoding** was applied with the first category dropped to mitigate multicollinearity risks. This process ensured the dataset was cleaned, transformed, and ready for robust modelling, contributing to reliable insights on Cholera risk factors.

# **4.0 Methodology**

This section outlines the methodology adopted for analysing and interpreting the risk factors of Cholera using a combination of statistical methods, interpretable machine learning (ML) models, and advanced model interpretability tools. The workflow involves three primary stages: machine learning model development, statistical analysis, and result aggregation.

Importantly, this data-driven analysis phase formed the foundation for the second component of this study: the development of an interactive LangChain-based LLM agent designed to synthesise findings, answer user queries, and provide policy-relevant explanations. The design and implementation of this agentic framework are presented in Section 5.

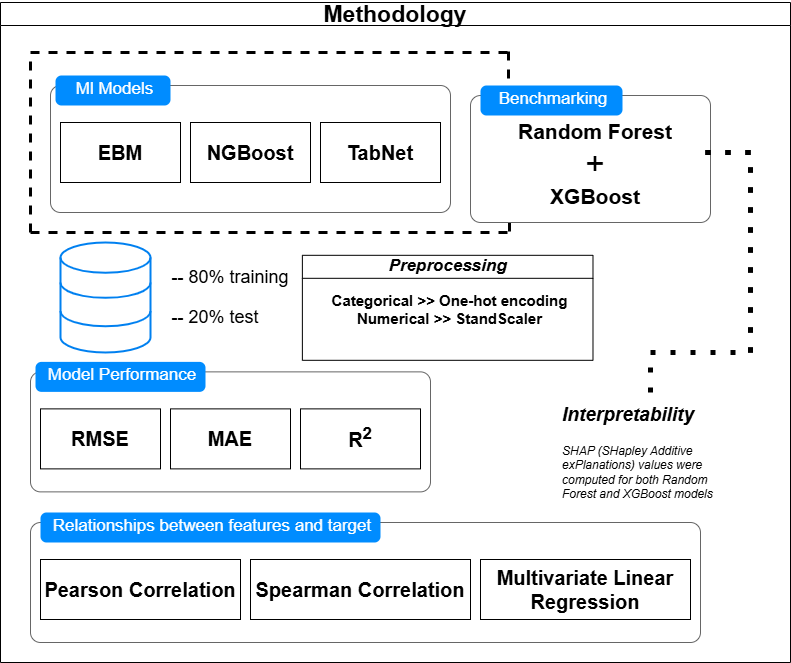


Figure Overview of the analytical workflow combining machine learning models, statistical validation

Figure 1 provides a visual summary of the analytical workflow adopted in this study to identify and interpret key risk factors associated with Cholera incidence. The diagram captures the sequential process beginning with data preprocessing, which includes cleaning, encoding, scaling, and transforming features to ensure compatibility with machine learning models. The flow then diverges into two parallel tracks: one for statistical analysis and the other for machine learning. On the statistical analysis side, techniques such as Pearson and Spearman correlation, as well as multivariate linear regression, are applied to examine both linear and monotonic relationships between predictors and the Cholera outcome variable. These help validate the importance of features and highlight statistically significant factors. On the machine learning side, models including EBM, NGBoost, TabNet, Random Forest, and XGBoost are trained on the processed dataset. Their performance is evaluated using standard regression metrics (RMSE, MAE, and R²). Model interpretability is enhanced using SHAP values, which provide both global and local explanations of feature influence on predictions.

## **4.1 Justification of Model Selection**

We selected Explainable Boosting Machine (EBM) due to its glass-box interpretability, allowing health policymakers to understand model reasoning through intuitive plots and feature scores. NGBoost was chosen for its probabilistic forecasting capabilities, providing not just point estimates but also confidence intervals—critical for public health uncertainty. TabNet, a modern deep learning model optimized for tabular data, was included for its sparse attention-based feature selection, which enables both high performance and built-in interpretability. Random Forest and XGBoost served as benchmarks, as they are widely used and understood in epidemiological machine learning applications.

Each CSV file generated from the analysis plays a targeted role in the framework. These files are directly fed into the LLM agent’s context to support traceable, data-grounded responses.

* feature\_importance\_comparison.csv: Contains rankings across models, showing which variables most influence Cholera incidence.
* model\_performance\_comparison.csv: Summarizes RMSE, MAE, and R² metrics to compare model efficacy.
* pearson\_corr\_matrix.csv and spearman\_corr\_matrix.csv: Capture linear and monotonic relationships between predictors and the outcome.
* linear\_regression\_coefficients\_pvalues.csv: Includes coefficients and p-values from multivariate regression, validating statistical significance.

To cross-validate machine learning interpretations, Pearson and Spearman correlations were used to identify consistent associations between predictors (e.g., water contamination levels, sanitation metrics) and Cholera incidence. These correlations served as initial sanity checks. A multivariate linear regression model was then applied to assess the statistical significance of each feature when controlling for others. This ensured that variables identified by ML were not only predictive but also causally plausible.

# **5.0 Agentic LLM Framework Design**

Following the interpretable machine learning and statistical analyses described in Section 4, we developed an interactive agentic system capable of responding to user queries about Cholera risk factors, model performance, and recommendations. This system is implemented using the LangChain and LangGraph frameworks and deployed via a Streamlit front-end, creating a usable interface for both researchers and decision-makers.

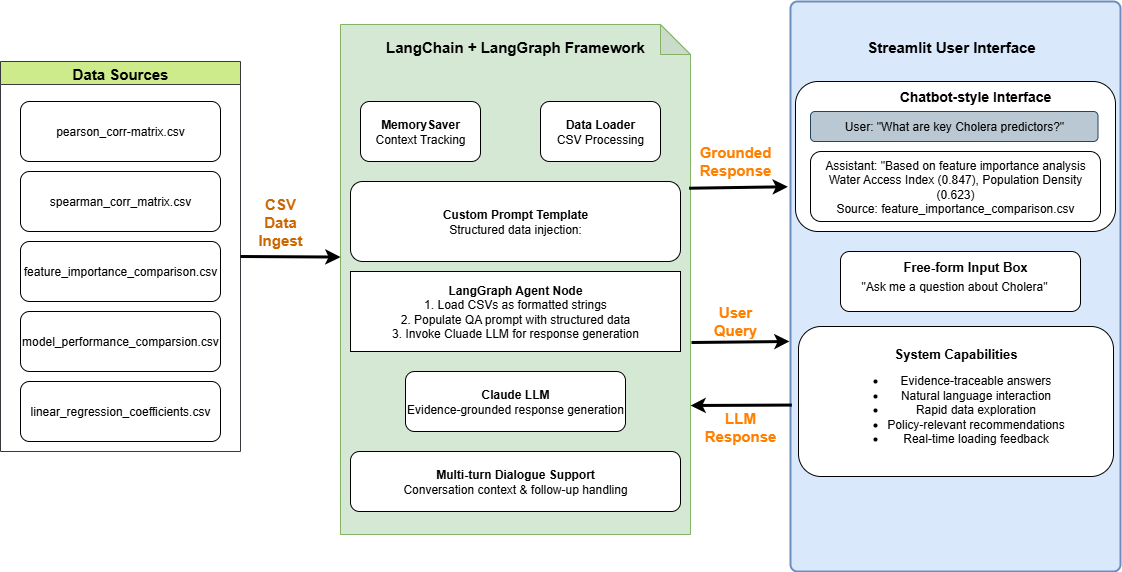


Figure Architecture of the Agentic LLM Framework for Cholera Risk Assessment

Figure 2 shows the system architecture of the Agentic LLM Framework for Cholera Risk Assessment, illustrating the end-to-end integration of machine learning outputs, statistical analysis, and a user-facing conversational interface. At the core of the architecture is a LangGraph agent powered by Claude LLM, designed to ingest structured CSV files containing feature importance rankings, correlation matrices, regression outputs, and model performance metrics. These files, generated from the earlier modelling phase, are accessed through LangChain's memory and utility functions to dynamically populate a prompt template that guides the language model's reasoning. The system leverages LangChain’s structured state handling and LangGraph’s workflow capabilities to manage the flow of user input, agent response, and contextual memory across dialogue turns.

The Streamlit interface provides a clean, interactive platform where users—whether researchers, policymakers, or public health officials—can pose epidemiological and policy-relevant questions using natural language. The framework supports evidence-traceable answers, where responses reference specific CSV-backed insights, such as "according to the Pearson correlation matrix…" or "as seen in the feature importance rankings from EBM…". The interface also includes visual feedback like loading indicators, conversation history, and persistent memory to enable multi-turn interactions. The diagram further depicts how the components are interconnected: CSV data feeds into the agent, which in turn communicates with the user interface bidirectionally. This hybrid design enhances accessibility for non-technical stakeholders while ensuring scientific robustness and reproducibility of insights.Top of Form

Bottom of Form

## **5.1 LangChain + LangGraph Integration**

The core reasoning component of the framework is constructed using LangChain's **LangGraph** abstraction, which enables stateful multi-step interactions via directed graphs. A minimal LangGraph was created consisting of a single node responsible for handling question-answering, supported by structured memory and prompt-driven generation. To ensure informed and context-aware responses, the agent loads the outputs of earlier ML and statistical stages. Each file is loaded into memory and injected into a custom prompt template, enabling the LLM (Claude) to generate grounded, evidence-supported responses. The prompt includes clearly delimited sections for correlation matrices, regression coefficients, and model importance rankings, facilitating targeted answer generation. The agent node is then registered as a step within the LangGraph workflow, and its behaviour is defined using the following logic:

* Load required CSVs as formatted string segments.
* Populate the QA prompt using the structured data.
* Append user input and prompt to the message history.
* Invoke the LLM (model\_claude) for response generation.

The agent supports dynamic, multi-turn dialogue, with memory persistence handled via MemorySaver, allowing coherent follow-ups and conversation context tracking.

## **5.2 Streamlit UI**

The interactive component of the framework is deployed through a lightweight **Streamlit** web application, making the system accessible to non-technical users. The interface includes:

* A chatbot-style layout with a persistent conversation history.
* An input box allowing free-form user questions (e.g., *“What are the most important features for predicting Cholera?”*).
* A dynamic loading spinner for response feedback.

Upon submitting a question, the user's message is added to the ongoing state, passed to the LangGraph agent, and the response is appended and rendered.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure Deployed Q&A interface showing traceable, data-driven responses to Cholera-related queries, with citations from correlation and feature importance analyses.

**Figure 3** presents a screenshot of the deployed Q&A interface, illustrating the interaction between user queries and LLM-generated responses grounded in structured data analysis. The interface displays structured data citations from correlation matrices and feature importance analyses, demonstrating the system's ability to provide traceable, data-driven epidemiological insights.

## **5.3 System Capabilities**

This integrated framework enables a robust and user-friendly interface for evidence-grounded exploration of Cholera risk factors and mitigation strategies. One of the core strengths of the system lies in its ability to produce evidence-traceable answers. Each response generated by the agent is not merely speculative or generic; rather, it is grounded in structured data inputs—including correlation matrices, regression outputs, and machine learning-derived feature importance scores—formatted as CSVs. These files are loaded and programmatically embedded within the prompt context supplied to the LLM, ensuring that the output references actual model results and statistical relationships. This traceability ensures that users can verify and interpret the rationale behind the answers, lending credibility and transparency to the system.

In addition, the framework supports natural language interaction, allowing users to pose questions in an intuitive, open-ended format. Whether a policymaker seeks insights into the top environmental contributors to Cholera outbreaks, or a health researcher wants to compare model performances, the system parses the query and retrieves relevant content without requiring the user to understand the technical structure of the underlying data. This makes the agent accessible to non-specialist stakeholders and promotes interdisciplinary usability.

Furthermore, the system supports rapidexploration of complex analytical outputs. By synthesising diverse statistical and machine learning results into a single LLM-driven pipeline, the agent can quickly summarise key predictors, explain uncertainty, and offer evidence-based recommendations. Users can ask follow-up questions in the same session, enabling dynamic and multi-faceted inquiry. For instance, after identifying that poor sanitation and high nitrate levels are key predictors, a user could ask how these variables vary across countries or how mitigation might reduce predicted Cholera incidence. The system responds fluidly, supported by memory persistence and structured prompting logic.

Overall, this hybrid design exemplifies a novel application of agentic LLMs in the epidemiological domain. By combining explainable machine learning, statistical validation, and conversational AI, the framework offers a powerful tool for data-driven public health insights. It provides a replicable, transparent, and interactive platform for guiding policy decisions, especially in resource-limited settings where timely and interpretable guidance is critical.

# **Results and Insights**

To complement the machine learning models and statistical correlation analysis, we conducted a Poisson regression analysis to further investigate the association between predictor variables and the incidence of Cholera. Poisson regression is particularly suitable for modeling count-based outcome variables, such as disease incidence rates, and allows for direct estimation of incidence rate ratios (IRRs) while controlling for multiple covariates simultaneously.

This approach enabled us to quantify the relative contribution of key environmental, socio-economic, and infrastructural predictors to Cholera cases per 100,000 people, while accounting for potential confounding effects. The model included all relevant predictors that passed prior significance thresholds in the correlation and regression analyses. Results from the Poisson regression are presented visually through three plots:

Figure 4 presents the results of the Poisson regression analysis using three complementary visualizations, each offering key insights into the environmental and socio-economic factors associated with cholera incidence. The first plot (A), a forest plot, reveals several variables with statistically significant positive associations with cholera cases. Among the most notable are the pH level of water, contaminant level (ppm), and lead concentration (µg/L), all of which point to water quality as a critical driver of cholera risk. Diarrheal cases per 100,000 people also emerge as a strong predictor, likely indicating overlapping infrastructural vulnerabilities that facilitate both cholera and other waterborne diseases. Additionally, higher ambient temperatures and elevated dissolved oxygen levels are associated with increased cholera incidence, aligning with the known environmental preferences of Vibrio cholerae.

The second plot (B), displaying incidence rate ratios (IRRs), confirms that increases in factors such as pH, turbidity, contaminant level, and temperature are consistently linked to elevated cholera risk. Though the IRRs are modest in magnitude, their statistical significance underscores the public health importance of these variables. Interestingly, urbanization rate also shows a slight positive association, suggesting that rapid urban growth—often accompanied by inadequate infrastructure—may contribute to increased disease vulnerability.

The third plot (C), a volcano plot, compares the statistical and practical significance of each variable's effect on cholera cases. While many variables are statistically significant, their effect sizes tend to be modest, falling within a narrow practical range. This indicates that cholera risk is not driven by a single dominant factor but rather by a network of interrelated influences. No variables exhibit both high statistical and practical significance, reinforcing the multifactorial nature of cholera transmission. Overall, these findings emphasize the need for integrated, multi-dimensional public health interventions that address a broad set of environmental and infrastructural conditions to effectively mitigate cholera risk.

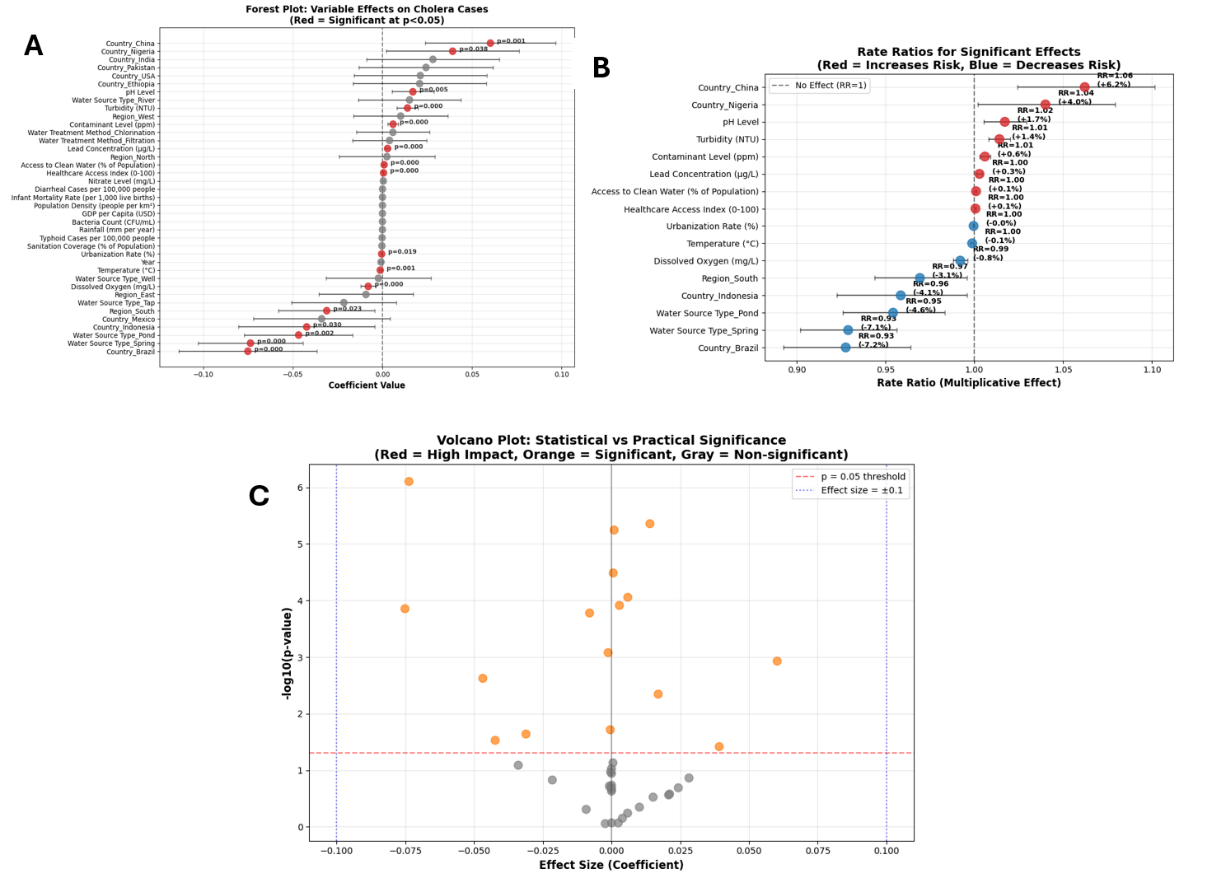
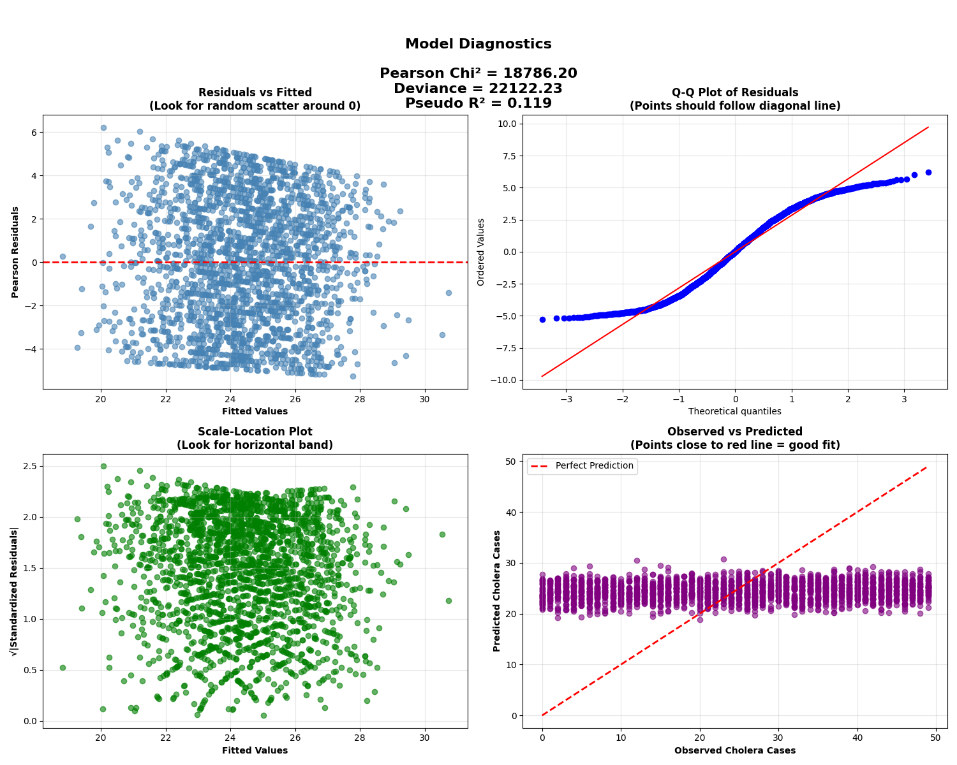


Figure Poisson regression analysis

Figure 5 below presents the model diagnostic plots for the Poisson regression predicting cholera cases, offering a comprehensive evaluation of model fit and underlying assumptions. The Pearson residuals vs. fitted values plot (top-left) is expected to show a random scatter of points around zero. However, the observed dispersion and slight funnel shape suggest potential heteroscedasticity, indicating that residual variance may increase with higher fitted values. This challenges the assumption of constant variance and suggests possible under- or overestimation in specific ranges. The Q-Q plot (top-right), which evaluates the normality of residuals, shows notable deviations from the diagonal red line, particularly in the tails. Although strict normality isn't required for Poisson models, this deviation may reflect unmodeled complexity, outliers, or limitations in the log-linear structure of the model. The scale-location plot (bottom-left), ideally displaying a horizontal spread of points, shows some variance across fitted values. This further supports concerns about heteroscedasticity and suggests that predictive performance may not be consistent across the outcome range.

The observed vs. predicted plot (bottom-right) compares actual and predicted cholera case counts. The clustering of points away from the 45-degree red reference line indicates systematic bias, with the model failing to capture the full variability in observed data. This is consistent with the model’s pseudo R² of 0.18, suggesting limited explanatory power. While the Poisson model identifies significant associations, Figure 5 highlights concerns related to residual behaviour, prediction accuracy, and variance explanation. These findings indicate the potential need for model enhancements, such as incorporating interaction terms, exploring nonlinear effects, or adopting alternative approaches like negative binomial regression to better account for overdispersion and improve overall model fit.



# Discussion

# Conclusion and Future Work

# References