ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 9 (Issue 1) (2025) Pages 21-30

A Hybrid Epidemiological Model Approach to Improvement of Predictive Accuracy in Zambian Infectious Diseases Modelling

Grey Chibawe[1], Mayumbo Nyirenda[2]

University of Zambia

1. gchibawe@gmail.com; 2. mayumbo@gmail.com@gmail.com

Abstract **- Recurrent infectious disease outbreaks, including cholera and influenza, as well as recent global pandemics like COVID-19, pose persistent public health challenges in Zambia. Traditional compartmental models based on Ordinary Differential Equations (ODEs), particularly Susceptible-Exposed-Infectious-Recovered (SEIR) frameworks, have long been used to predict disease spread. While these models are relatively simple and require fewer data, they often lack the flexibility to capture non-linear and stochastic factors—such as environmental variables and abrupt policy shifts—that can critically influence epidemic trajectories in resource-limited settings. In contrast, Artificial Neural Network (ANN) approaches excel at learning complex, non-linear relationships directly from data. By incorporating diverse inputs (e.g., climatic variables, demographic distributions), ANNs can adapt to evolving outbreak patterns more effectively than traditional ODE-based methods. However, their reliance on large, high-quality datasets and considerable computational resources can hinder adoption in places with fragmented surveillance systems. To address these complementary strengths and weaknesses, this study explores a hybrid modelling strategy that integrates a parameter-optimised SEIR model with a Transformer-based ANN. Historical COVID-19 data from 2020 to 2024 and environmental data (temperature, rainfall, humidity) were used to develop and validate three models: (1) an SEIR model whose parameters were estimated via curve fitting, (2) a standalone Transformer ANN, and (3) a combined SEIR-ANN ensemble. Model performance was assessed using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). Results indicate that the hybrid model consistently outperformed the individual SEIR and ANN models, exhibiting the lowest RMSE and MAE. Furthermore, integrating environmental factors into the ANN substantially improved predictive accuracy. These findings highlight the promise of hybrid frameworks in capturing the multifaceted dynamics of infectious diseases in Zambia. By leveraging SEIR's mechanistic insights alongside the ANN's capacity to learn from diverse datasets, public health practitioners can improve outbreak predictions and resource allocation. Nevertheless, barriers—such as limited data availability, computational infrastructure, and model interpretability—must be addressed to foster broader implementation. Strengthened data collection systems, increased investment in computational tools, and targeted capacity-building programs are recommended to fully realise the benefits of hybrid epidemiological modelling in Zambia.**

Keywords: Epidemiological Modelling, SEIR Model, Integral-based SEIR, Artificial Neural Networks (ANN), Predictive Accuracy, COVID-19.

I. Introduction

The recurrence of disease epidemics has become a significant public health challenge in Zambia, with outbreaks occurring almost annually. These outbreaks range from pandemics, such as COVID-19, to endemic diseases that persist in specific regions of the country. Infectious diseases like cholera, which often result in high mortality rates during outbreaks, contrast with diseases like influenza, which, while less fatal, may exacerbate symptoms and complications due to co-infections within vulnerable populations. Epidemiologists and public health officials have been employing mathematical models to understand and predict the dynamics of these diseases, aiming to mitigate their spread and impact before reaching peak prevalence [1] [2].

Traditionally, compartmental models based on Ordinary Differential Equations (ODEs), such as the Susceptible-Exposed-Infectious-Recovered (SEIR) framework, have been widely used in Zambia for modelling infectious disease dynamics. These models divide populations into distinct compartments and rely on parameters such as infection rates, recovery rates, and population sizes. However, traditional SEIR models often fail to capture the non-linear complexities of real-world disease spread, including environmental, socioeconomic, and behavioural factors [3]. To address some of these limitations, integral models have been introduced, incorporating control dynamics such as lockdowns and environmental variables. Despite these improvements, traditional ODE-based models have shown limited predictive accuracy, particularly for smaller populations and regions [4].

The emergence of Artificial Neural Network (ANN)-based models has opened new avenues for modelling infectious disease dynamics. Unlike ODE models, ANN approaches are data-driven and capable of learning complex, non-linear relationships from large datasets. This capability allows ANN models to integrate diverse variables, such as climatic, demographic, and behavioural factors, into their predictions [5]. However, ANN models also have limitations, such as requiring extensive training datasets, which may not always be available in resource-limited settings like Zambia [6] [7].

In Zambia, hybrid modelling approaches have been explored to combine the strengths of traditional compartmental models and ANN-based methods. These hybrid models aim to leverage the parameter-based structure of ODE models while incorporating the flexibility and non-linear predictive capabilities of ANNs. Despite these efforts, existing hybrid models have often prioritised data mining over accurate prediction of disease dynamics, leaving room for improvement in their predictive performance [8], [7], [9].

This study conducts a comparative analysis of the predictive accuracy of traditional SEIR models and ANN-based models for infectious diseases in Zambia. By evaluating these models' performance and exploring hybridisation approaches, the study aims to propose an optimised modelling framework for improved prediction accuracy and public health intervention planning.

A. *Research Objectives*

The main objective of this study was to enhance the predictive accuracy and applicability of epidemiological models in Zambia by evaluating the effectiveness of Artificial Neural Networks (ANNs) and hybrid SEIR-ANN models, identifying their advantages and limitations, and proposing strategies to address key challenges such as data availability, computational infrastructure, and model interpretability. The specific objectives were:

1) To evaluate the effectiveness of Artificial Neural Networks (ANNs) in modelling and predicting infectious diseases in Zambia.

2) To analyze the performance of hybrid SEIR-ANN models in improving the predictive accuracy of epidemiological models compared to traditional SEIR frameworks.

3) To identify barriers to the adoption and optimization of ANN and hybrid models for epidemiological applications in Zambia, with a focus on data availability, computational infrastructure, and interpretability.

B. *Research Questions*

1) How effective are Artificial Neural Networks (ANNs) in predicting infectious diseases in Zambia compared to traditional SEIR models?

2) What are the advantages and limitations of hybrid SEIR-ANN models in capturing disease dynamics and improving predictive accuracy?

3) What are the key barriers to adopting ANN and hybrid models for epidemiological modelling in Zambia, and how can they be addressed to enhance their utility?

C. *Research Hypotheses*

H1: Hybrid SEIR-ANN models provide significantly higher predictive accuracy in forecasting infectious disease outbreaks in Zambia compared to standalone SEIR or ANN models.

H2: Incorporating environmental factors such as temperature, humidity and rainfall into ANN-based epidemiological models improves the models' predictive accuracy in comparison to models that exclude such variables.

II. Literature Review

A. *Traditional SEIR ODE Models in Epidemiological Modelling in Zambia*

Mathematical modeling has been a cornerstone of epidemiology, providing insights into the spread and control of infectious diseases. Among these models, the Susceptible-Exposed-Infectious-Recovered (SEIR) framework, based on Ordinary Differential Equations (ODEs), has been extensively used for its simplicity and adaptability in compartmentalizing disease dynamics [10]. In Zambia, the SEIR model has played a crucial role in understanding outbreaks of diseases such as cholera, malaria, and influenza, where its deterministic nature has allowed policymakers to forecast the trajectory of epidemics under various scenarios [11]. Equations 1 to 4 are the four S, E, I and R ODE equations for the SEIR compartmentalised model.

Equation (1) represents the rate of change of the susceptible population into being exposed to the infectious disease. The term β is the transmission rate, I is the infectious population, and N is the total population (N=S+E+I+R).

$$\frac{dS}{dt} = -\beta \frac{SI}{N} \tag{1}$$

Equation (2) models the rate of change of the exposed population into becoming infectious. The parameter σ is the rate at which exposed individuals become infectious.

$$\frac{dE}{dt} = \beta \frac{SI}{N} - \sigma E \tag{2}$$

Equation (3) models the rate of change of the infectious population into recovery or being removed from the population. γ is the rate at which individuals recover from the infectious disease.

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{3}$$

Equation (4) represents the rate of change of the recovered population. Individuals recover and gain immunity but can also be susceptible again.

$$\frac{dR}{dt} = \gamma I \tag{4}$$

One significant advantage of the SEIR model lies in its ability to provide a structured framework for disease dynamics. By dividing the population into distinct compartments (e.g., susceptible, exposed, infectious, and recovered), the model captures transitions between states based on defined rates such as infection and recovery. This compartmentalization has been particularly useful in Zambia, where health systems often lack the capacity for real-time surveillance. With limited data, SEIR models can leverage parameters derived from historical outbreaks to estimate critical thresholds for interventions like vaccination or quarantine [12]. Moreover, its simplicity enables quick implementation and analysis, making it accessible to researchers and policymakers with limited computational resources [13] [14].

Despite its utility, the applicability of SEIR models in Zambia has faced several challenges. One major shortcoming is the assumption of homogeneity within compartments, which overlooks individual variations such as age, gender, and comorbidities that can significantly influence disease transmission and recovery rates. This limitation is particularly critical in Zambia, where demographic diversity and varying socioeconomic conditions play a pivotal role in disease dynamics [15]. Additionally, SEIR models assume constant transmission rates, which may not reflect the reality of fluctuating environmental and behavioral factors, such as seasonal rainfall patterns that exacerbate cholera outbreaks or variations in public health interventions [15] [16].

Another notable limitation is the model's deterministic nature, which fails to account for stochastic elements inherent in disease transmission, particularly in small populations or during the early stages of an outbreak. In Zambia, where rural communities often experience localized epidemics, these stochastic effects can lead to significant deviations from model predictions, reducing their reliability [17] [4]. Furthermore, SEIR models struggle to incorporate complex interactions between diseases. For instance, co-infections such as HIV and tuberculosis, which are prevalent in Zambia, create dynamic feedback loops that are challenging to model within the traditional SEIR framework [18] [19].

Efforts to address these shortcomings have led to the development of extended SEIR models that incorporate additional compartments or parameters. For example, models have been adapted to include environmental variables such as water quality and temperature, which are critical for cholera dynamics. However, these extensions often require extensive data inputs, which are not always available in resource-constrained settings like Zambia [20]. Hybrid approaches, combining SEIR models with data-driven techniques such as machine learning, are also being explored to improve predictive accuracy and adaptability. While promising, these approaches necessitate robust datasets and computational infrastructure, presenting another layer of complexity [21].

In summary, while traditional SEIR models have been instrumental in epidemiological modeling in Zambia, their limitations underscore the need for more nuanced and adaptable approaches. The deterministic and compartmental nature of these models provides a foundation but requires enhancements to address the complexities of real-world disease dynamics. Recent advancements, including hybrid and extended models, offer pathways to overcome these limitations, though their implementation in Zambia remains contingent on data availability and computational capacity.

B. Integral-Based SEIR Models in Epidemiological Modelling in Zambia

Integral-based SEIR models represent a significant advancement in epidemiological modelling, particularly in addressing some of the limitations inherent in traditional ODE-based frameworks. By incorporating integral equations, these models capture time-dependent factors and dynamic interventions, such as lockdowns, vaccination campaigns, and behavioural changes, providing a more realistic representation of disease dynamics [4]. In Zambia, where epidemic management often involves abrupt policy shifts and environmental changes, integral-based models have shown potential in offering more accurate and flexible predictions.

One of the primary advantages of integral-based SEIR models is their ability to account for variable transmission rates. Unlike traditional SEIR models, which assume constant rates, integral models allow for the incorporation of time-varying parameters, such as seasonal variations and the impact of public health interventions. For instance, during cholera outbreaks in Zambia, rainfall patterns and water sanitation efforts significantly influence disease transmission. Integral-based models can integrate these temporal factors, enabling more precise predictions of outbreak peaks and durations [4].

Furthermore, integral models provide a framework for incorporating memory effects, where the current state of the epidemic depends on its historical trajectory. This feature is particularly relevant in Zambia, where previous outbreaks of diseases like malaria and cholera influence current dynamics through acquired immunity and residual environmental contamination. By considering these historical dependencies, integral-based models offer a more comprehensive understanding of disease progression and recovery [16] [22].

Despite their advantages, the implementation of integral-based SEIR models in Zambia faces several challenges. One major limitation is the increased computational complexity associated with solving integral equations, which often require numerical methods and significant computational resources. In resource-constrained settings, this complexity can limit their accessibility and adoption by local researchers and public health officials [23]. Additionally, integral models demand high-quality, time-resolved data to parameterise and validate the models accurately. In Zambia, where data collection systems are often fragmented and incomplete, this requirement poses a significant barrier to effective implementation [24].

Another limitation is the difficulty in integrating multi-disease interactions within integral-based frameworks. While these models excel in capturing temporal dynamics, they struggle to represent the complex interdependencies between co-circulating diseases. For example, the interactions between HIV, tuberculosis, and COVID-19 in Zambia create synergistic effects that are challenging to model using existing integral-based approaches. Addressing these interactions requires

hybrid models that combine the strengths of integral equations with data-driven techniques, such as machine learning, to enhance predictive accuracy and adaptability [25] [26].

Recent advancements in computational tools and data availability offer promising avenues for overcoming these challenges. The integration of satellite-derived environmental data and mobile health technologies has the potential to improve the temporal resolution of input data, enhancing the applicability of integral-based models in Zambia. Moreover, collaborations between local and international research institutions can provide the computational infrastructure and expertise needed to develop and deploy these models effectively [27].

In conclusion, integral-based SEIR models offer a powerful extension to traditional ODE frameworks, addressing critical limitations related to temporal dynamics and intervention modelling. While their application in Zambia is still in its early stages, ongoing advancements in data collection and computational resources hold promise for their broader adoption. By leveraging these models, public health officials can gain deeper insights into the dynamics of infectious diseases, enabling more effective interventions and policy decisions.

C. Current ANN and Hybrid Epidemiological Models in Zambia and Their Potential Improvement

Artificial Neural Networks (ANNs) and hybrid models have emerged as powerful tools for epidemiological modelling, leveraging their ability to handle complex, nonlinear relationships and large datasets. In Zambia, the application of these models has gained traction in recent years, particularly in the wake of the COVID-19 pandemic, which highlighted the limitations of traditional compartmental models in capturing real-world disease dynamics [23] [28] [29] [30].

ANNs excel in scenarios where traditional models struggle, such as incorporating environmental, demographic, and behavioural factors into disease prediction. For instance, studies have utilised ANNs to predict cholera outbreaks by integrating rainfall, temperature, and water quality data, demonstrating superior accuracy compared to SEIR models [31] [32] [33]. These models are also adept at handling incomplete or noisy data, a common challenge in Zambia's resource-constrained settings, by employing advanced techniques such as data imputation and feature selection [34].

Hybrid models, which combine the mechanistic foundations of traditional SEIR or integral-based models with the predictive capabilities of machine learning algorithms, offer a promising avenue for epidemiological modelling in Zambia. These models utilise the compartmental structure of SEIR frameworks to simulate disease dynamics while leveraging machine learning to refine parameter estimates and improve predictive accuracy. For example, a hybrid model incorporating SEIR dynamics and a neural network has been used to predict malaria outbreaks in Zambia, achieving enhanced performance by accounting for both disease transmission mechanisms and environmental drivers [35] [36].

Despite their advantages, the adoption of ANN and hybrid models in Zambia faces several barriers. One significant challenge is the lack of high-quality, large-scale datasets required to train these models effectively. While traditional SEIR models can operate with limited data, ANNs rely on extensive historical data to capture complex patterns and relationships. In Zambia, fragmented health information systems and inconsistent data reporting pose significant obstacles to the development of these models [37].

Moreover, the computational demands of ANN and hybrid models can be prohibitive in resource-limited settings. Training these models often requires advanced hardware and software infrastructure, which may not be readily available in many regions of Zambia. This limitation underscores the need for capacity-building initiatives and investments in computational resources to enable the broader adoption of these models [6].

Another challenge lies in the interpretability of ANN models. While they excel at prediction, their "black-box" nature can make it difficult for public health officials to understand the underlying factors driving model outputs and thereby adopt their usage. Hybrid models address this issue to some extent by combining the transparent structure of SEIR frameworks with the predictive power of ANNs, but further efforts are needed to enhance the explainability of these models to ensure their utility in decision-making processes [38] [39].

To improve the applicability and effectiveness of ANN and hybrid models in Zambia, several strategies can be pursued. First, efforts should focus on strengthening data collection and integration systems to provide the high-quality datasets required for model development. Initiatives such as mobile health platforms and satellite-based environmental monitoring can enhance data availability and resolution. Second, investments in computational infrastructure and capacity-building programs for local researchers and public health officials are crucial to enable the development and implementation of these models. Finally, research should prioritise the development of interpretable hybrid models that balance predictive power with transparency, ensuring their outputs are actionable for policymakers.

II. Methodology

To address the research objectives, quantitative experimental approach was employed, leveraging on secondary data:

A. Data Collection and Preparation

1) Historical epidemiological data for COVID-19 were obtained from the CDC international database. Historical climatic data spanning the period 2019 to 2023 were obtained from the online repository Visualcrossing.com. These data included weekly infection case numbers of diseases and environmental factors (e.g., temperature, rainfall and humidity). For this study, only COVID-19 data were used in the experimentation.

2) Data preprocessing involved cleaning, normalization, and handling missing values to ensure high-quality datasets suitable for model development. Weeks with missing values were replaced with zeros for mathematical reasons. While COVID-19 data were originally available as daily counts, they

were aggregated into weekly data to maintain consistency with datasets for other diseases, such as cholera, which are recorded on a weekly basis because this phase of the research was part of an ongoing research to integrate multiple diseases in disease outbreak forecasting. Although this resampling ensured temporal alignment across diseases in the broader multi-disease forecasting study, it introduced potential smoothing effects that may have masked short-term outbreak fluctuations or rapid transmission dynamics. Furthermore, inconsistencies in daily reporting may have carried over into the aggregated weekly counts. Weeks with incomplete data were filled with zero values to maintain continuity for SEIR model simulation, but this may have biased the model toward underestimating incidence rates in low-reporting periods. While normalisation and trend-based validation methods were applied to improve data consistency, the absence of metadata on age, gender, and co-morbidities restricted the development of stratified or subgroup-specific forecasts.

B. Model Development

1) Traditional SEIR Models: Baseline SEIR and integral-based SEIR models were implemented using differential equations to simulate disease dynamics. This was done using Python code in the SPYDER IDE.

2) ANN Models: Artificial Neural Networks were developed using historical data. Input features included environmental factors to capture nonlinear relationships.

3) Hybrid Models: SEIR frameworks were combined with ANN components to refine parameter estimates and enhance predictive accuracy. For example, ANNs were used to predict time-varying transmission rates, which were then fed into SEIR models.

4) Environmental Factors: The prediction accuracy of the ANN model was evaluated, comparing prediction when environmental factors (temperature, humidity and rainfall) where not integrated in the model and when they were integrated in the model.

C. Model Validation and Comparison

1) Models were validated using a test dataset (2023 data) and evaluated based on predictive accuracy metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) values.

2) Sensitivity analyses were conducted to assess the robustness of the models under different scenarios, such as varying environmental conditions or intervention measures.

D. Implementation Framework

A framework for implementing hybrid models in Zambia's public health system was developed, including recommendations for capacity-building, data integration, and policy support.

E. Software and Tools

Python in the SPYDER IDE was used for model development and analysis. Libraries such as TensorFlow, PyTorch, and SciPy facilitated ANN and SEIR model implementation.

F. Computational Environment and Constraints

The models were implemented using Python (TensorFlow, SciPy) in the Spyder IDE on a standard consumer-grade laptop with 16GB RAM and a Quad-Core Intel Core i7 processor (2.6GHz). Due to hardware limitations, model training—particularly for the Transformer-based ANN—was capped at a modest number of epochs (up to 200), and batch sizes were optimised to prevent memory overflow. More complex ANN architectures or extensive hyper-parameter tuning could not be fully explored due to these computational constraints. Despite this, results achieved stable convergence and acceptable performance. Future implementations could benefit from GPU acceleration or cloud-based training platforms for improved efficiency and scalability.

G. Hypotheses Testing

The research hypotheses were addressed through a structured experimental framework that involved the development and evaluation of three distinct models using historical COVID-19 data and environmental variables from Zambia. To test H1, the study compared the predictive performance of three models: (1) a traditional SEIR model enhanced with parameter estimation via curve fitting, (2) a Transformer-based Artificial Neural Network (ANN), and (3) a hybrid SEIR-ANN model that combined mechanistic modelling with machine learning. To evaluate H2, two ANN models were trained and tested—one without environmental variables and another with temperature, rainfall, and humidity as additional input features.

II. Results

A. Data Collection and Preparation

This research adopted an experimental quantitative non-participant approach to evaluate the efficacy of a Transformer neural network in improving the prediction accuracy of COVID-19 mathematical epidemiological models in Zambia. The study utilized COVID-19 daily infections data collected from the CDC, CIDRZ and WHO websites. Ethical approval was obtained from the University of Zambia National Science Research Ethics Committee (Appendix 1) and the National Health Research Authority (Appendix 2) to ensure data privacy and integrity.

This study took an experimental quantitative approach. The data used ranged from week 1 of the year 2020 and week 35 of the year 2024. This study was done concurrently with a study on Cholera whose data were weekly, so, the COVID-19 daily infections data had to be converted to weekly to match the Cholera data. Eighty percent of the data were used to train the model while twenty percent were used to test the model.

B. SEIR Model with Parameter Estimation

The Baseline SEIR model is usually simplistic. Parameters are predetermined and hard-coded. Parameters (especially beta) have to be manually edited to observe change in the shape of the compartmental curves. In this research, curve fitting was utilised to find the best fit beta which could estimate as close as possible predicted daily infections which matched the actual data. In the model, an initial beta of 0.25 was guessed which was used in the method snippet below (Fig. 1) to curve fit so that a best fit beta value was found which brought the prediction

curve as close as possible to the actual data. From the model, new infections were computed using equation 5.

$$new\ infections = \sigma \times E(t) \quad (5)$$

```
##############################################################
#########################
# 5. FIT BETA TO THE WEEKLY DATA
##############################################################
#########################

        # Initial guess for beta
        initial_guess = [0.25]

# Curve fit the model_weekly(...) to
actual_weekly_cases_2023
        popt, pcov = curve_fit(model_weekly,
        weeks_2023, actual_weekly_cases_2023,
        p0=initial_guess)
        best_beta = popt[0]
        print("Best-fit beta =", best_beta)
```

Fig. 1 Except of Python code from the Spyder IDE: Showing the method for curve fitting in the SEIR Baseline model.

The model was simulated and visualized using Python code in the Spyder IDE. Fig. 2 is the plot from the output of the equation 5 of the model. We called the model the SEIR Model with Parameter Estimation because of deriving beta from curve fitting. The best fit beta was 0.09970113105348326. We calculated the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE) and the R-squared (R^2) for the model as 593.138, 216.002 and -0.456 respectively.

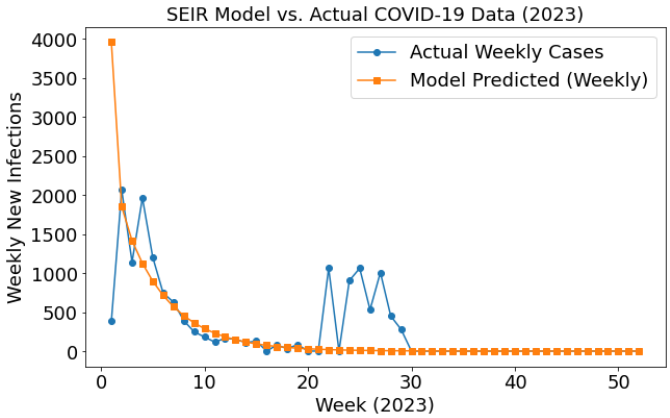


Fig. 2 Predicted weekly infections vs Actual data weekly infections of COVID-19 in the SEIR Model with Parameter Estimation.

C. Basic Artificial Neural Network Prediction of Weekly COVID-19 Infections

We developed a basic Artificial Neural Network and ran it using four algorithms namely, Convolutional Neural Network (CNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) and Transformer in order to choose the one that would outperform the others as the algorithm to use in the ensemble or hybrid model. As seen in Table 1, the Transformer outperformed the other three; its RMSE was lowest. This was true whether at lower number of epochs or higher number of epochs. The ANNs were created using TensorFlow in Python. Fig. 3 is the visualisation output of the models.

TABLE I
Accuracy Comparison of Four ANN Algorithms on COVID-19 Prediction of Zambian 2023 Weekly Infections, Using Metrics RMSE, MAE and R^2

S/No.	Algorithm	RMSE	MAE	R^2
1	CNN	447.95	397.75	0.17
2	GRU	387.76	315.91	0.38
3	LSTM	379.21	264.96	0.40
4	Transformer	375.47	185.38	0.42

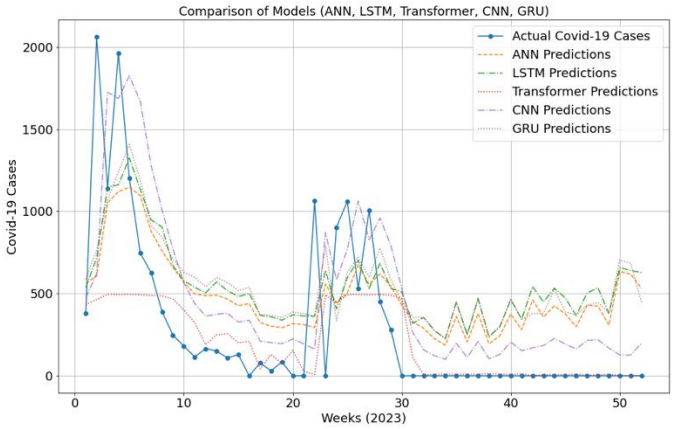


Fig. 3 Comparison of prediction accuracy of four ANN models in the prediction of COVID-19 based on Zambian disease surveillance data.

D. The SEIR/ANN Ensemble Model

We decided to create a hybrid (or ensemble) model to leverage on the improved accuracy of both the SEIR by Parameter Estimation and the ability of ANN to learn complex systems from data. The model begins by fitting the SEIR to find its best parameters (e.g., β) and generates weekly predictions. The generated weekly predictions are used as a weekly explanatory feature when training the Transformer model. In the Python code, the final hybrid model is evaluated on test data (2023 COVID-19 data) to check whether the prediction accuracy is better than those of either the SEIR by Parameter Estimation or the Transformer ANN. Table 2 shows the comparison of the RMSE, MAE and R^2 of the three models. Fig. 4 is the visualisation of the three models: the SEIR by Parameter Estimation, the Transformer ANN and the Hybrid model.

TABLE 2
Comparison of Prediction Accuracy of SEIR Model Alone, Transformer Model Alone and the SEIR + Transformer Hybrid Model Using RMSE, MAE And R^2 in the Prediction of COVID-19 Based on Zambian Data

S/No.	Model	RMSE	MAE	R^2
1	SEIR by Parameter	593.138	216.002	-0.456

	Estimation Model *alone*			
2	Transformer ANN Model *alone*	392.228	218.765	0.363
3	SEIR + Transformer Hybrid Model	387.699	190.060	0.378

This research utilised the outcome of the accuracy metrics (RMSE, MAE and R^2) to evaluate prediction accuracy. Future research could test the statistical significance of the observed differences in predictive accuracy by carrying out statistical tests such as a paired t-test to compare prediction errors between the models. The inclusion of environmental data potentially excluded over-fitting of the models to just the COVID-19 data making the outcome of this research generalisable to other infectious diseases prevalent in Zambia.

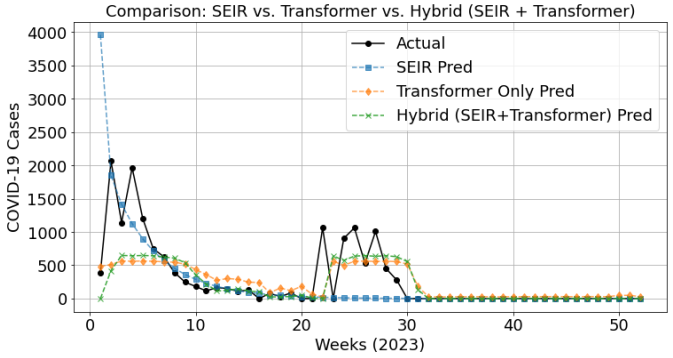


Fig. 4 Visualisation of the SEIR by Parameter Estimation alone, Transformer Model alone and the SEIR + Transformer Hybrid Model in the prediction of Zambia's COVID-19.

E. The SEIR/ANN Ensemble Model

We evaluated the prediction accuracy of the Transformer ANN to see if there was improvement when additional data such as environmental factors were integrated in the model. Fig. 5 shows that the RMSE and MAE were lower (Table 3) when environmental factors (temperature, humidity and rainfall) were integrated in the model. Fig. 6 and Fig. 7 are plots of the model without and with environmental factors respectively. Line plot slopes for the prediction were more realistic in Figure 6 when environmental factors had been included.

E. Hypotheses Testing

To test H1, the study compared the predictive performance of the three models: (1) a traditional SEIR model enhanced with parameter estimation via curve fitting, (2) a Transformer-based Artificial Neural Network (ANN), and (3) a hybrid SEIR-ANN model that combined mechanistic modelling with machine learning. The hybrid model consistently exhibited lower RMSE and MAE values (Table 2), indicating superior accuracy, thus supporting H1. The evaluation of H2 was done through the training of two ANN models and their testing—one without environmental variables and another with temperature, rainfall, and humidity as additional input features. The inclusion of these environmental factors resulted in a notable decrease in RMSE and MAE, as well as a higher R^2 value (Table 3), demonstrating improved predictive performance and thus validating H2.

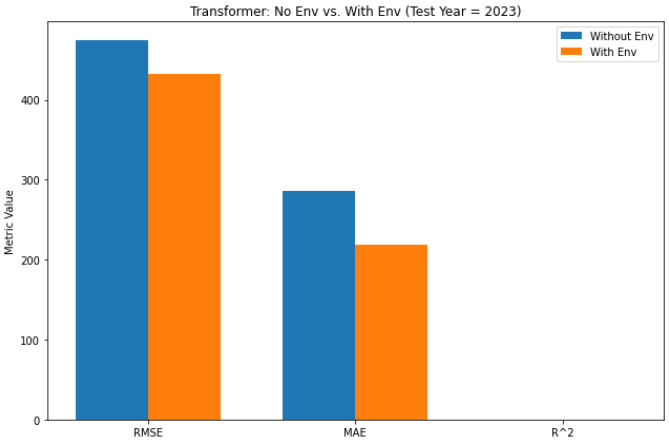


Fig. 5 Comparison of the Transformer ANN model without and with environmental factors in the modelling of the Zambian COVID-19 outbreak.

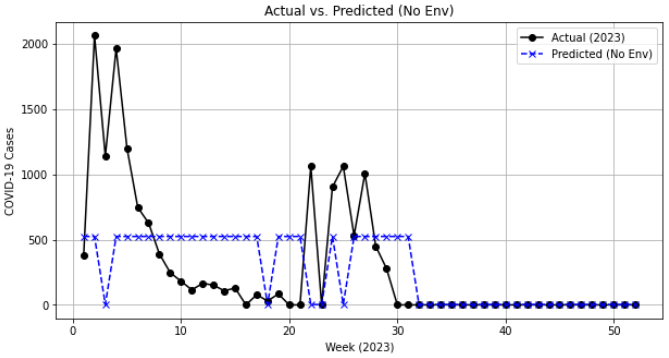


Fig. 6 Line plot of the Transformer ANN model when environmental factors were not included in the modelling of the Zambian COVID-19 outbreak.

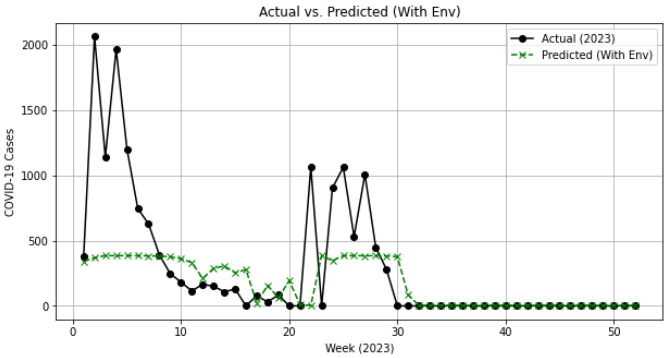


Fig. 7 Line plot of the Transformer ANN model when environmental factors were included in the modelling of the Zambian COVID-19 outbreak.

TABLE 3

Comparison of Prediction Accuracy, Through RMSE, MAE And R^2, of the Transformer ANN in the Prediction of Zambia's COVID-19 Outbreak

Constituents of Transformer ANN Model	**RMSE**	**MAE**	**R^2**
No Environmental Factors	474.299	285.816	0.069
Environmental Factors Included	432.034	218.854	0.227

II. Discussion and Conclusion

ODEs compartmentalised SEIR models' simplicity make them easily accessible to epidemiological researchers. They can easily be used where data is scanty. However, they fail to integrate complex stochastic features of disease dynamics like effect of environmental factors and interventions (e.g. lockdowns) [13] [14] [15] [16]. This study improved the baseline SEIR model by leveraging on the computer's ability to process fast. The SEIR by Parameter Estimation model was developed which begins with fitting the SEIR curve to estimate the best fit beta whose plot will closely match the actual data plot. In the model, the initial beta value is guessed which is used in the curve fitting to select the best fit beta value. So, whatever value is guessed would result in the same best fit output. This improved prediction accuracy of the model as seen in Figure 1 where the negative slope closely matches the actual data. The results of the SEIR by Parameter Estimation were compared with those of the selected ANN algorithm (Transformer). ANN are data dependent. They can learn complex systems to almost match the patterns in the data. ANN can replicate stochastic situation better than ODEs SEIR models. Bringing the advantage of the SEIR by Parameter Estimation and that of the ANN together, we developed the SEIR + Transformer Hybrid model and compared its accuracy with both the improved SEIR model alone and the Transformer model alone. Table 3 shows that the hybrid model was superior; its RMSE and MAE were lower than those of the two models alone. This showed improved accuracy when ordinary SEIR models and ANN models are ensembled.

One of the limitations of ANN models is their "black-box" nature, which reduces interpretability and may hinder adoption by public health practitioners. To address this, future iterations of the hybrid model could incorporate explainability techniques such as SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME). These tools can help identify the most influential features contributing to predictions—such as rainfall, temperature, or prior infection levels—thereby enhancing transparency and trust in model outputs. By revealing the contribution of each input variable, these methods can also guide data collection priorities and public health interventions more effectively.

Artificial Neural Networks are data driven [40]. The more data they are fed with the more accurately they learn. Figure 4 and Table 3 showed the Transformer ANN model's improvement in prediction accuracy when the environmental factors were included. The RMSE, MAE and R^2 were lower after integrating environmental factors in the model. The lower the error metrics the more accurate a model is [41].

The results of this study are clear evidence that if epidemiologists leverage on the computing power of computers to ensemble models, their predictions would be better accurate and quicker to resolve epidemics. Accuracy improves more when ordinary mathematical models and ANN models are ensembled relying on computing. We, therefore, recommend that epidemiologists leverage on the power of the computer and ensemble models into hybrid models, especially in African countries like Zambia where data collection is not consistent. While data-driven neural networks have the potential to capture complex, stochastic dynamics, their greater computational resources, demand for larger datasets and specialised expertise may hinder the adoption of the neural networks in real-time epidemiological settings. Simpler compartmental models are more widely utilised due to their relative transparency with what happens to the data, lighter computational demands, and well-established theoretical foundations [42]. The results of this research can be implemented in a web application which epidemiological research could use to resolve disease outbreaks better. This could improve adoption of neural networks-based models. We also recommend that stakeholders and government in Zambia should channel enough resources (financial and material) towards consistent data collection during disease epidemics to help epidemiologists produce more accurate models.

References

[1] S. B. R. H. M. A.-A. M. C. G.G. Kolaye, "Mathematical assessment of the role of environmental factors on the dynamical transmission of cholera," *Communications in Nonlinear Science and Numerical Simulation*, vol. 67, pp. 203-222 https://doi.org/10.1016/j.cnsns.2018.06.023, 2019.

[2] R. M. Mullner, epidemiology. Encyclopedia Britannica, 2024.

[3] P. Kumari, "Mathematical Modelling of Malaria Transmission Dynamics in Rural Regions of Jharkhand and Bihar: An SEIR Model Approach," *medicine*, vol. 6, no. 2, pp. 137-140, 2024.

[4] S. D. K. &. T. G. K. Z. Fodor, "Why integral equations should be used instead of differential equations to describe the dynamics of epidemics," *arXiv preprint*, vol. arXiv:2004.07208, 2020.

[5] H. R. M. &. G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications," *Environmental modelling & software*, vol. 15, no. 1, pp. 101-124, 2000.

[6] M. K. D. M. D. L. M. A. D. M. E. N. K. H. K. R. M. G. J. K. &. H. M. J. M. Tshimula, "Artificial Intelligence for Public Health Surveillance in Africa: Applications and Opportunities," *arXiv preprint*, vol. arXiv:2408.02575, 2024.

[7] P. K. R. H. M. L. M. F. C. P. M. &. A. R. C. Mulenga, "Predicting mortality in hospitalized COVID-19 patients in Zambia: an application of machine learning," *Global Health, Epidemiology and Genomics*, vol. 2023, no. e8, 2023.

[8] J. S. &. D. Kunda, "Challenges of using Data Mining Techniques to Analyze and Forecast COVID-19 Pandemic in Zambia," *Zambia ICT Journal*, vol. 6, no. 1, pp. 7-17, 2022.

[9] M. C. C. C. V. C. F. L. &. D. K. J. Kalezhi, "Modelling Covid-19 infections in Zambia using data mining techniques," *Results in Engineering*, vol. 13, no. 100363, 2022.

[10] M. J. K. &. P. Rohani, Keeling, M. J., & Rohani, P. (2008). Modeling infectious diseases in humans and animals, Princeton university press, 2008.

[11] J. Sichone, "Estimating the basic reproduction number for the 2015 Nyimba district bubonic plague outbreak (Doctoral dissertation, The University of Zambia).," University of Zambia, Lusaka, 2018.

[12] J. H. &. J. S. Morris, "Infectious Disease Modeling," *Annual Review of Statistics and Its Application*, vol. 12, 2024.

[13] J. T. &. T. Luong, "Modeling epidemics with compartmental models," *Jama*, vol. 323, no. 24, pp. 2515-2516, 2020.

[14] C. C. K. W. W. &. R. M. S. J. Panovska-Griffiths, "Mathematical modeling as a tool for policy decision making: Applications to the

COVID-19 pandemic," *In Handbook of Statistics*, vol. 44, pp. 291-326, 2021.

[15] N. K. F. D. M. C. N. H. N. C. S. P. B. P. K. M. B. N. D. N. &. A. C. C. Y. Yang, "Challenges addressing inequalities in measles vaccine coverage in Zambia through a Measles–Rubella supplementary immunization activity during the COVID-19 pandemic," *Vaccines*, vol. 11, no. 3, p. 608, 2023.

[16] K. D. B. Y. J. A. H. R. R. C. A. J. M. Usmani, "A Review of the Environmental Trigger and Transmission Components for Prediction of Cholera," *Tropical medicine and infectious disease*, vol. 6, no. 3, p. 147, 2021.

[17] Plan, Z. M. C. E., "Plan, Zambia Multisectoral Cholera Elimination," Plan 2019-2025, Lusaka, Zambia, 2028.

[18] J. H. P. S. K. Z. F. K. S.-S. M. M. C. A. G. R. E. R. A. P. P. E. T. &. C. G. S. G. Loevinsohn, "Respiratory pathogen diversity and co-infections in rural Zambia," *International Journal of Infectious Diseases*, vol. 102, pp. 291-298, 2021.

[19] T. C. &. C. L. C. Katamba, "HIV, syphilis and hepatitis B coinfections in Mkushi, Zambia: a cross-sectional study," *F1000Research*, vol. 8, p. 562, 2020.

[20] S. A. S. H. C. C. E. H. T. S. M. K. C. M. E. F. O. &. G. Z. K. A. Assamagan, "A study of COVID-19 data from African countries," *arXiv preprint*, vol. arXiv:2007.10927, 2020.

[21] D. J. R. P. &. J. C. T. S. Chen, "Leveraging advances in data-driven deep learning methods for hybrid epidemic modeling," *Epidemics*, vol. 48, p. 100782, 2024.

[22] T. K. A. M. J. B. X. R. E. B. &. M. P. Baracchini, "Seasonality in cholera dynamics: A rainfall-driven model explains the wide range of patterns in endemic areas," *Advances in water resources*, vol. 108, pp. 357-366, 2017.

[23] C. D., ""Comprehending symmetry in epidemiology: A review of analytical methods and insights from models of COVID-19, Ebola, Dengue, and Monkeypox,"," *Medicine*, vol. 2, pp. 27-54, 2024.

[24] T. F. Menkir, "Equitable Infectious Disease Modeling for Data-Constrained Settings and Data-Overlooked Populations," (Doctoral dissertation), 2024.

[25] A. H. &. Z. A. J. H. Jones, "Transmission-dynamics models for the SARS Coronavirus-2," *American Journal of Human Biology*, vol. 32, no. 5, 2020.

[26] G. Gibson, "APPLIED INFECTIOUS DISEASE FORECASTING FOR PUBLIC HEALTH," 2021.

[27] A. Alam, "Cloud-based e-learning: development of conceptual model for adaptive e-learning ecosystem based on cloud computing infrastructure," in *International Conference on Artificial Intelligence and Data Science*, 2021, December.

[28] F. Baldo, Informed machine learning for epidemics: from data analysis to time-series forecasting, 2024.

[29] B. &. T. C. I., "Integrating Agent-Based and Compartmental Models for Infectious Disease Modeling: A Novel Hybrid Approach," *arXiv preprint*, vol. arXiv:2407.20993, 2024.

[30] S. &. M. M. S. B., "A new approach to the dynamic modeling of an infectious disease," *Modelling of Natural Phenomena*, 2021.

[31] M. B. A. J. D. J. C. L. G. N. E. O. &. D. H. G. Luque Fernández, "Influence of temperature and rainfall on the evolution of cholera epidemics in Lusaka, Zambia, 2003–2006: analysis of a time series," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 103, no. 2, pp. 137-143, 2009.

[32] M. D., An Exploratory Study on the Opportunities and Challenges of using Machine Learning in the DHIS2 Ecosystem, 2023.

[33] O. E. O. O. T. C. O. &. S. O. O. H., "Data-Driven Machine Learning Techniques for the Prediction of Cholera Outbreak in West Africa," *Western European Journal of Modern Experiments and Scientific Methods*, vol. 1, no. 1, pp. 33-51, 2023.

[34] B. A., Targeted machine learning approaches for leveraging data in resource-constrained settings, Berkeley: University of California, 2020.

[35] P. N. I. K. E. L. K. B. M. T. &. N. B. K. E. K., "Forecasting malaria morbidity to 2036 based on geo-climatic factors in the Democratic Republic of Congo," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12271, 2022.

[36] F. B. F., "A LOCAL-LEVEL SPATIAL MODEL OF MALARIA TRANSMISSION SITES IN RURAL ZAMBIA," *(Doctoral dissertation, Johns Hopkins University)*, 2018.

[37] C. Z., "A review of the successes and challenges of coordination and collaboration in the implementation of e-government programmes: a case of Zambia," 2021.

[38] A. A. M. P. G. F. F. A. R. V.-K. U. R. T. R. &. P. M. A. S. F., "Covid-19 outbreak prediction with machine learning," *Algorithms*, vol. 13, no. 10, p. 249, 2020.

[39] N. J. F. T. M. H. L. &. Y. C. H. P. H., "A hybrid model with spherical Fuzzy-AHP, PLS-SEM and ANN to predict vaccination intention against COVID-19," *Mathematics*, vol. 9, no. 23, p. 3075, 2021.

[40] S. Haykin, Neural networks: a comprehensive foundation, PTR: Prentice Hall, 1998.

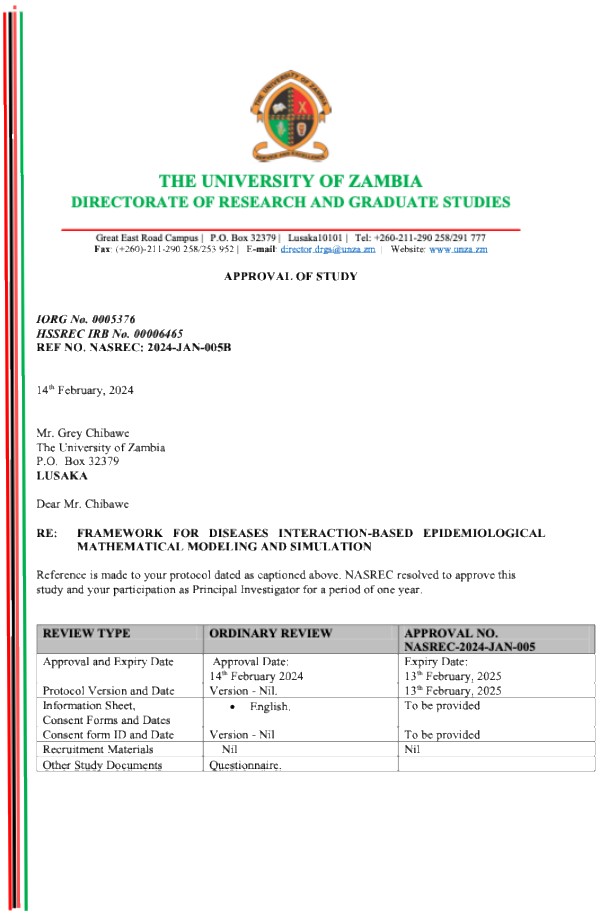
[41] R. J. Hyndman, Forecasting: principles and practice, 2018.

[42] A. H. &. G. Katriel, "Mathematical modelling and prediction in infectious disease epidemiology," *Clinical microbiology and infection*, vol. 19, no. 11, pp. 999-1005, 2013.

APPENDICES

APPENDIX I

Ethical Clearance by the UNZA Natural Sciences Research Ethical Committee (NASREC)



THE UNIVERSITY OF ZAMBIA
DIRECTORATE OF RESEARCH AND GRADUATE STUDIES

Great East Road Campus | P.O. Box 32379 | Lusaka10101 | Tel: +260-211-290 258/291 777
Fax: (+260)-211-290 258/253 952 | E-mail: director.drgs@unza.zm | Website: www.unza.zm

APPROVAL OF STUDY

IORG No. 0005376
HSSREC IRB No. 00006465
REF NO. NASREC: 2024-JAN-005B

14th February, 2024

Mr. Grey Chibawe
The University of Zambia
P.O. Box 32379
LUSAKA

Dear Mr. Chibawe

RE: FRAMEWORK FOR DISEASES INTERACTION-BASED EPIDEMIOLOGICAL MATHEMATICAL MODELING AND SIMULATION

Reference is made to your protocol dated as captioned above. NASREC resolved to approve this study and your participation as Principal Investigator for a period of one year.

REVIEW TYPE	ORDINARY REVIEW	APPROVAL NO. NASREC-2024-JAN-005
Approval and Expiry Date	Approval Date: 14th February 2024	Expiry Date: 13th February, 2025
Protocol Version and Date	Version - Nil.	13th February, 2025
Information Sheet, Consent Forms and Dates	• English.	To be provided
Consent form ID and Date	Version - Nil	To be provided
Recruitment Materials	Nil	Nil
Other Study Documents	Questionnaire.	

APPENDIX II

Ethical Clearance by the Zambia National Health Research Authority (NHRA)

NATIONAL HEALTH RESEARCH AUTHORITY

Lot No. 18961/M, off Kasama Road, Chalala, P.O. Box 30075, LUSAKA
Tell: +260211 250309 | Email: znhrasec@nhra.org.zm | www.nhra.org.zm

NHRA-1369/12/07/2024 **23rd July 2024**

The Principal Investigator,
Mr. Grey Chibawe,
The University of Zambia P.O. Box 32379 LUSAKA,
Lusaka.

Dear Mr. Grey Chibawe,

Re: Request for Authority to Conduct Research

The National Health Research Authority Is in Receipt of Your Request for Authority to Conduct Research Titled **"FRAMEWORK FOR DISEASES INTERACTION-BASED EPIDEMIOLOGICAL MATHEMATICAL MODELING AND SIMULATION"**

I wish to inform you that following submission of your request to the Authority, our review of the same and in view of the ethical clearance, this study has been **approved** on condition that:

1. The relevant Provincial and District Medical Officers where the study is being conducted are fully appraised.
2. Progress updates are provided to NHRA bi-annually from the date of commencement of the study.
3. The final study report is cleared by the NHRA before any publication or dissemination within or outside the country.
4. After clearance for publication or dissemination by the NHRA, the final study report is shared with all relevant Provincial and District Directors of Health where the study was being conducted, University leadership, and all key respondents.

Yours sincerely,

National Health Research Authority

Prof. Victor Chalwe,
Acting Director/Chief Executive Officer