LEVERAGING GEO-NLP FOR ENHANCED ANTIRETROVIRAL DRUG DISTRIBUTION IN NIGERIA: INSIGHTS FROM SOCIAL MEDIA AND NEWS DATA

Bashirudeen Opeyemi Ibrahim, Anthony Soronnadi, Olubayo Adekanmbi,
Ahmad Ibrahim Ismail & **David Akanji**
Research and Innovation Department
Data Scientist Networks
33, Queens Street, Yaba, Lagos, Nigeria
{bashirudeen, olubayo, ahmad, Anthony, david}@datasciencenigeria.ai

ABSTRACT

Faced with over 1.9 million HIV/AIDS cases, Nigeria's need for efficient antiretroviral therapy (ART) distribution is critical. Conventional assessment methods, restrained by logistical issues and data scarcity, require innovative solutions. This study employs Geographic Natural Language Processing (Geo-NLP) to analyse social media and news content, offering novel insights into public discourse on HIV/AIDS and ART across Nigeria. Using a custom Named-Entity Recognition (NER) model to process data from NairaLand and major newspapers, the research uncovers geographical patterns in HIV/AIDS-related conversations, achieving a significant model performance with an overall F1-Score of 83.27%. The findings highlight areas with intense discussions on HIV/AIDS, suggesting urban centres like Bauchi, Jos, and Ibadan as priority sites for targeted ART interventions. This approach promises to refine ART distribution strategies and sets a precedent for employing Geo-NLP in public health planning. Despite its brevity, the study underscores the potential of integrating Geo-NLP with traditional data to enhance healthcare delivery in Nigeria, paving the way for more effective public health interventions against the HIV/AIDS epidemic.

1 INTRODUCTION

As Nigeria is faced with the ongoing challenge of HIV/AIDS concerning its public health space, the need for innovative and novel solutions to improve antiretroviral drug distribution has never been more crucial. With over 1.9 million individuals living with HIV in Nigeria National Agency for the Control of AIDS (NACA) (2021), the efficient delivery of antiretroviral therapy (ART) stands as a cornerstone in the battle against this epidemic. Conventional approaches used in the evaluation of drug distribution necessities have often fallen short, limited by logistical complexities and the scarcity of localised data on HIV/AIDS prevalence and treatment needs. Olutuase et al., (2022) This research presents a pioneering approach by employing Geographic Natural Language Processing (Geo-NLP) to analyse conversations on social media and in news publications, aiming to uncover insights into the public's perception and the spatial distribution of discussions related to HIV/AIDS and antiretroviral drugs across Nigeria. The prevalence of digital platforms and the proliferation of news media offer an untapped reservoir of data, reflecting the concerns, needs, and gaps in the current healthcare delivery system. By developing a specialised Named-Entity Recognition (NER) model, this study parses vast amounts of unstructured text data from NairaLand, Nigeria's largest online forum, and various leading newspapers. The model's ability to identify and categorise relevant entities provides a nuanced understanding of the geographical variances in HIV/AIDS-related discussions and drug mentions. This methodological innovation enhances the granularity of data available to healthcare stakeholders and opens new paths for targeted interventions. The anticipated outcome of this research is a comprehensive map that highlights regions with pronounced discussions on HIV/AIDS and antiretroviral drugs, indicating potential hotspots for intervention. Such a data-driven approach promises to revolutionise the planning and implementation of antiretrovi-

ral drug distribution strategies, making them more receptive to the actual needs and conversations happening within communities across Nigeria. In doing so, this research aligns with the global endeavour to end the HIV/AIDS epidemic, contributing valuable insights towards achieving the United Nations' Sustainable Development Goal of good health and well-being for all.

2 Related Works

2.1 Geo-NLP in Health Interventions

Geo-NLP fuses the abilities of Natural Language Processing (NLP) and Geographic Information Systems (GIS) to extract and scrutinise geographically anchored data from textual sources. This intersection of disciplines has shown increasing value in health-related specialisations, enabling improved pharmacovigilance, public health surveillance, and the optimisation of health services. In an illustrative study by Nikfarjam et al. (2015), the significance of NLP for identifying adverse drug reactions within social media narratives was showcased, highlighting the importance of leveraging user-generated content for pharmacovigilance objectives. Similarly, research conducted by Gammino et al. (2014) employed GIS technologies to scrutinise the efficacy of vaccination campaigns in Northern Nigeria, showing the profound impact of geospatial analytics in improving the management and delivery of healthcare services.

2.2 Social Media as a Data Source for Public Health

Conway et al. (2019) examine the utilisation of NLP in analysing social media for public health research, emphasising its role in detecting health-related trends and beliefs. This review emphasises the shift from infectious disease monitoring to mental health and substance abuse topics, indicating the expanding scope of Geo-NLP applications in public health. Calvo et al. (2017) explore the use of NLP in inferring mental states from social media posts, providing a pathway for using Geo-NLP to understand and address mental health issues at a population level.

3 Research Methodology

3.1 Data Collection and Preparation

3.1.1 Data Source

Our research utilised a comprehensive dataset from NairaLand, a popular online forum in Nigeria. We specifically focused on discussions and posts on antiretroviral (ARV) medications, as listed in the 7th Edition of the Nigerian Essential Medicines List published by the Federal Ministry of Health (2020). We also included relevant articles from prominent Nigerian newspapers, ThisDay and Punch, to capture a broader perspective on HIV/AIDS discourse in the country.

3.1.2 Data Cleaning and Integrity

The dataset, comprising approximately 160,000 words, underwent rigorous cleaning to ensure the quality and integrity of the information. This process involved the removal of special characters and anonymising personally identifiable information to maintain privacy standards and the reliability of the dataset. The cleaning procedures were meticulously designed to preserve the contextual accuracy and relevance of the data.

3.1.3 Data Utilisation

The initial dataset's subset of 52,858 words was meticulously selected for labelling and developing the Named-Entity Recognition (NER) model. This selection was based on criteria aimed at optimising the quality and relevance of the data for our specific analytical goals.

3.2 Data Labelling

3.2.1 Labelling Tool

We employed the NER Text Annotator, a free, web-based tool designed explicitly for categorising text data into predefined entity tags for the labelling process. This tool's user-friendly interface and compatibility with SpaCy facilitated efficient and accurate labelling.

3.2.2 Entity Tags

The data was categorised into six primary entity tags:

Table 1: Entity Tags and Descriptions

S/N	Entity Tag	Description
1	HIV/AIDS-Related Terms	Terms and phrases specifically related to HIV/AIDS, including medical terminology and interventions.
2	Antiretroviral Drugs	Names and types of drugs used in the treatment and management of HIV/AIDS.
3	Geographical Location	Specific locations mentioned in the text are relevant to the distribution and impact of HIV/AIDS.
4	Organisations	Names and details of organisations involved in HIV/AIDS treatment, research, or policy-making.

3.3 Named-Entity Recognition Training

3.3.1 Model Development

We utilised SpaCy2, a sophisticated and open-source natural language processing library, to construct a Named Entity Recognition (NER) model. The powerful capabilities of SpaCy2 were instrumental in developing a model capable of precisely identifying and classifying specified entity tags. The original JSON dataset was partitioned into three subsets for training, testing, and evaluation purposes, comprising 70%, 20%, and 10% of the data. This structured approach facilitated an efficient training process and thorough performance assessment of the NER model.

3.3.2 Model Selection and Evaluation

Model Training

The model was trained using an annotated JSON file over 172 epochs. We meticulously monitored and assessed the model's performance throughout this process, allowing for continuous refinement and optimization. This approach's iterative nature ensured the development of an accurate and efficient.

The training was conducted in a GPU-enabled environment, significantly enhancing our model training process's computation speed and efficiency.

Our training pipeline comprised two key components: 'tok2vec', responsible for converting tokens (words) into meaningful vectors (numerical representations), and 'ner', the component tasked with recognizing and categorizing named entities within the text. We initiated our training with a learning rate of 0.001, a critical parameter influencing the rate at which our model learns from the training data.

We rigorously monitored several performance metrics throughout the training to evaluate our model's effectiveness and accuracy. These included:

- **Losses:** We tracked the *LOSS TOK2VEC* and *LOSS NER* to gauge the model's prediction error. Continuous adjustments were made to minimize these values, enhancing the model's predictive accuracy.

- **F1 Score (ENTS_F):** Serving as a harmonized measure of precision and recall, the F1 score became a crucial indicator of our model's accuracy in identifying named entities correctly.
- **Precision (ENTS_P):** This metric helped us understand the accuracy of the model's positive predictions, ensuring that the entities identified were truly relevant.
- **Recall (ENTS_R):** By measuring the model's ability to capture all relevant entities, recall ensured our model's comprehensiveness in entity recognition.

Throughout the model's training, which included multiple epochs, we observed a consistent improvement in our model's performance. Concurrently, the F1 score, precision, and recall exhibited substantial growth, reflecting the model's enhanced ability to identify named entities accurately. The training culminated in a model achieving an F1 score of 87.01%, with corresponding precision and recall rates that underscored its high level of accuracy and reliability for our NER tasks.

Model Selection

Following the comprehensive training phase, we selected the best-performing model based on its superior F1 score, precision, and recall metrics. This model stood out for its exceptional balance between accurately identifying named entities and minimizing false positives, making it the ideal candidate for deployment in our subsequent analysis.

Model Evaluation

The best-performing model underwent evaluation on a designated 20% subset of the evaluation dataset, attaining an F1 score of 83.27%. This metric underscores the model's robustness in accurately identifying named entities, reflecting a well-balanced precision and recall.

Deployment and Analysis

The best-performing model was deployed to analyze a previously untouched subset of our dataset, comprising approximately 105,000 words. Applying our model to fresh data allowed us to gain new insights and validate the model's effectiveness in real-world scenarios. The analysis conducted with this labelled data provided a fresh perspective, enhancing our understanding of the dataset and demonstrating the model's practical utility in extracting meaningful information from natural language text.

3.4 Exploratory Data Analysis

3.4.1 Data Visualisation and Analysis Techniques

We employed various techniques to analyse and visually represent the labelled data:

- **Bar Charts:** These were used to illustrate the frequency and distribution of the top entities within each label.
- **Word Clouds:** Customized word clouds were generated for each label, visually representing the prevalence of each entity, with more frequent terms appearing larger.

3.4.2 Specific Analytical Focus

We conducted detailed analyses for the categories of 'GEOGRAPHICAL LOCATION', 'ORGANISATIONS', 'ANTIRETROVIRAL DRUGS', and 'HIV/AIDS-RELATED TERMS'. This included generating specific word clouds and bar charts to represent the data visually and quantitatively within these categories.

3.4.3 Geographical Emphasis

An emphasis was placed on analysing the 'GEOGRAPHICAL LOCATION' category, focusing on identifying and listing specific locations within Nigeria. This analysis aimed to understand the geographical distribution and impact of HIV/AIDS in the country

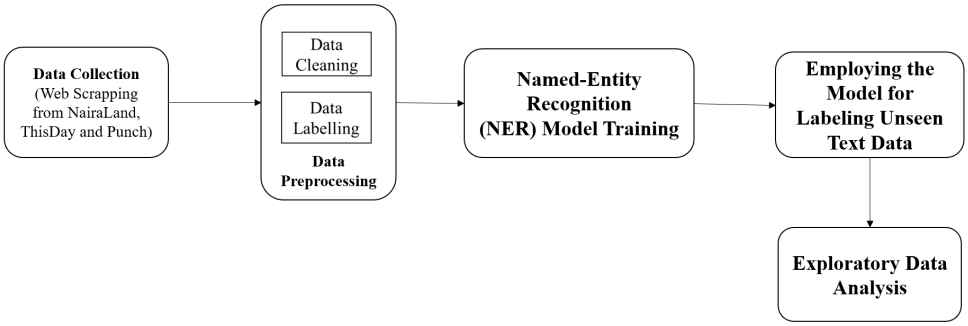

Figure 1: Flowchart of the Research Methodology

4 RESULT

4.0.1 MODEL EVALUATION

The table below shows the evaluation metrics based on the NER model trained using the spaCy framework. We obtained an overall F1-Score of 83.27.

S/N	Label	Precision	Recall	F1-Score
1	GEOGRAPHICAL LOCATION	89.04	81.93	85.34
2	ORGANISATIONS	86.90	82.95	84.88
3	HIV/AIDS-RELATED TERMS	85.11	80.48	82.73
4	ANTIRETROVIRAL DRUGS	83.55	79.87	81.67
	NER Overall	85.68	80.99	83.27

Table 2: NER Model Performance Metrics

4.1 EXPLORATORY DATA ANALYSIS

4.1.1 LABEL FREQUENCY DISTRIBUTION

The chart on the left displays the dataset's frequency distribution of different labels. The labels represent categories or classifications assigned to various entities within the dataset. The 'HIV/AIDS-RELATED TERMS' category has the highest count, indicating that it is the most common classification among the provided entities. This suggests a significant focus on HIV/AIDS-related terms within the dataset. The second most common label is 'ANTIRETROVIRAL DRUGS', followed by 'ORGANISATION' and 'GEOGRAPHICAL LOCATION'. This shows many entities related to medications for HIV/AIDS treatment, organisational references, and geographic entities.

4.1.2 TOP 20 ENTITIES

The right-hand chart highlights the top 20 entities across all categories, sorted by frequency in descending order. 'HIV' is the most frequently mentioned entity, aligning with the observation from the first chart that 'HIV/AIDS-RELATED TERMS' is a dominant label in the dataset. Following 'HIV', entities like 'World Health Organisation' and 'Nigeria' are among the most common. These likely relate to the reporting and discussion of HIV/AIDS within the contexts of global health and geographical focus (Nigeria). The presence of entities such as 'PEP' (post-exposure prophylaxis), 'tenofovir', 'lamivudine', 'efavirenz', 'lopinavir', and 'DTG' (dolutegravir) indicates specific antiretroviral medications or treatments that are frequently referenced, which corresponds to

	Entity	Label
0	Abuja	GEOGRAPHICAL LOCATION
1	National Assembly	ORGANISATIONS
2	National Assembly Committee on Finance Ministe...	ORGANISATIONS
3	National Assembly	ORGANISATIONS
4	National Assembly	ORGANISATIONS
...	...	...
1864	KATSACA	ORGANISATIONS
1865	HIV	HIV/AIDS-RELATED TERMS
1866	mothertochild transmission of HIV	HIV/AIDS-RELATED TERMS
1867	positive for the virus	HIV/AIDS-RELATED TERMS
1868	HIVAIDS	HIV/AIDS-RELATED TERMS

1869 rows × 2 columns

Figure 2: Prediction view of the NER on the Unseen Text Data

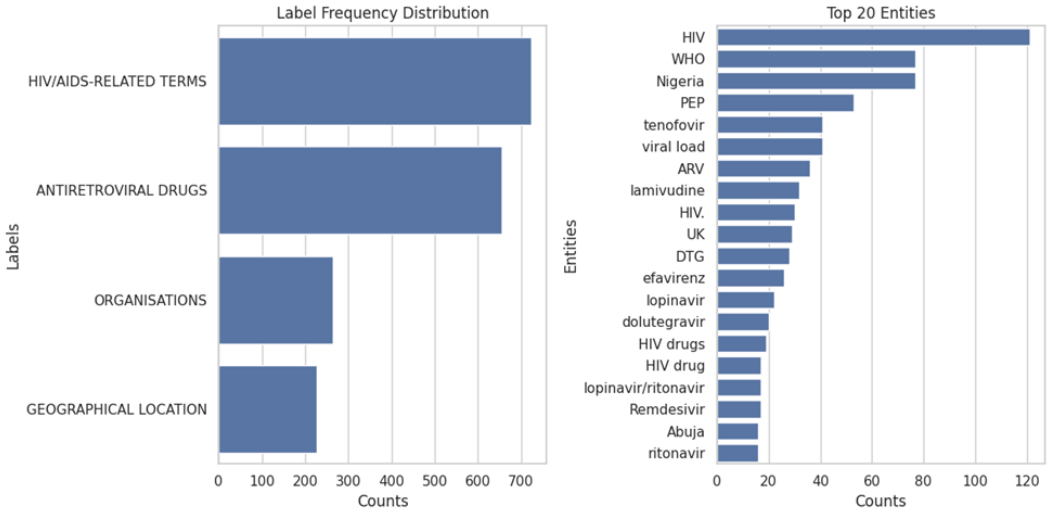

Figure 3: Bar chart showing the distribution of Label Frequency and Top 20 Entitiesn

the high label frequency of 'ANTIRETROVIRAL DRUGS'. Other entities like 'ARV' (antiretrovirals), 'HIV', 'ritonavir', 'UK', 'ARV', 'viral load', and 'dolutegravir' further emphasise the medical and geographical aspects of HIV/AIDS as focal points.

The most notable ARVs in the word cloud are lopinavir, ritonavir, tenofovir, and efavirenz. These are some of this category's most used and efficacious drugs. They belong to different classes of ARVs, such as protease inhibitors, nucleoside reverse transcriptase inhibitors, and non-nucleoside reverse transcriptase inhibitors. The most common ARV regimen in Nigeria was Tenofovir/Lamivudine/Dolutegravir, used by 77.4% of persons receiving PEPFAR-supported ART. Dirlikov et al., (2021). Other phrases in the word cloud, such as "reverse transcriptase inhibitors" and

Figure 4: Word Cloud for Entities Labelled as “ANTIRETROVIRAL DRUGS”

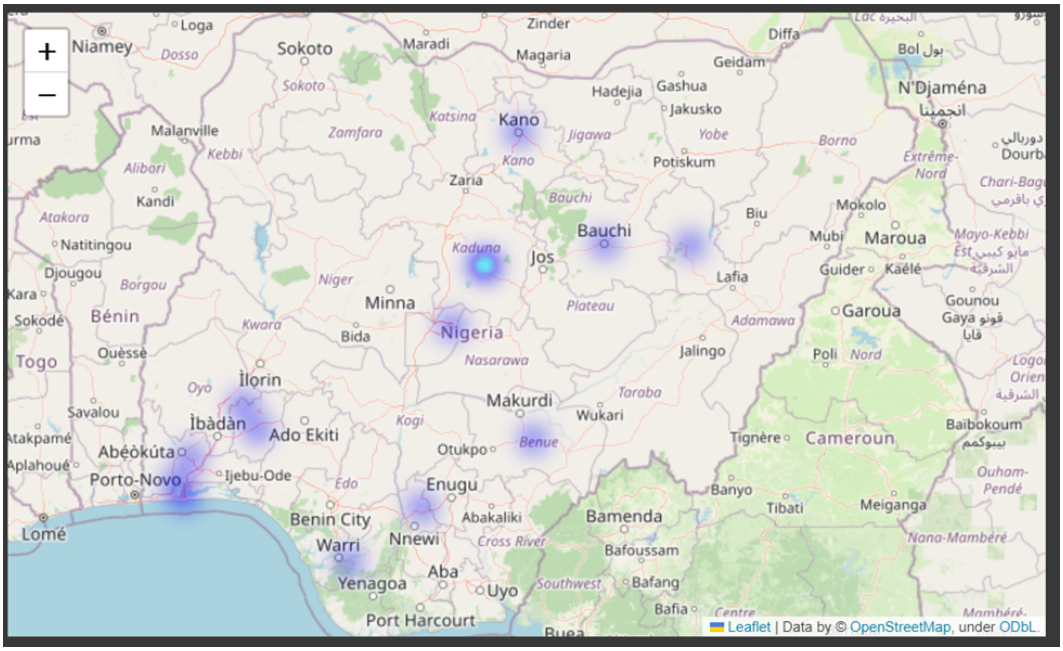

Figure 5: Heat Map showing the occurrence of Locations where the conversations around HIV/AIDS and Antiretroviral therapy have been had most in Nigeria

“nucleoside”, suggest these drugs’ classes or mechanisms of action. Reverse transcriptase inhibitors block an enzyme HIV needs to make copies of itself, while nucleoside analogues are building blocks of DNA that intrude with the viral replication process. These terms indicate the diversity and complexity of the ARV drugs and their modes of action.

The provided heat map, which overlaid the geographical map of Nigeria, was interpreted to reveal several critical points of interest. It was noted that there were significant concentrations of data points in central Nigeria, especially around Bauchi and Jos. This was taken to indicate a higher occurrence or density of the subject being studied, which could be factors such as HIV/AIDS case numbers, the distribution of healthcare facilities, or the distribution of ART. Further observations were made of discernible concentrations in the southwestern region near Ibadan, north of Lagos and

towards the southeast near Enugu. These areas were suggested to represent urban centres or other areas of heightened interest pertinent to the study's focus. Conversely, it was pointed out that the northern regions of Nigeria showed relatively less activity according to the heat map, suggesting either a lower relevance to the study or a lack of data points from those areas. The heat map was described as a visual representation that could significantly enhance understanding of geographical distributions and patterns. It was deemed particularly valuable for identifying potential hotspots for targeted interventions or further study within the context of public health and resource distribution. The necessity of correlating such visual data with on-the-ground realities and additional data sources was also emphasised to ensure accurate and comprehensive analysis.

5 DISCUSSION

5.1 MODEL PERFORMANCE AND ITS IMPLICATIONS

The overall F1-Score 83.27 achieved by the NER model indicates strong performance in accurately identifying and classifying entities across the predefined categories. The precision and recall metrics across specific categories, such as 'GEOGRAPHICAL LOCATION', 'ORGANISATIONS', 'HIV/AIDS-RELATED TERMS', and 'ANTIRETROVIRAL DRUGS', reflect the model's ability to handle the nuances of natural language in diverse contexts. This high level of accuracy is critical for ensuring the reliability of the extracted data, which forms the basis for subsequent analysis and insights.

5.1.1 IMPLICATIONS OF LABEL FREQUENCY DISTRIBUTION

The predominance of 'HIV/AIDS-Related Terms' and 'Antiretroviral Drugs' in the label frequency distribution indicates a high level of public attention to these topics on social media and in news publications. This engagement shows general concern and awareness about HIV/AIDS and the medications used, indicating a possible readiness of the population to participate in and support health intervention programs.

5.1.2 SIGNIFICANCE OF TOP 20 ENTITIES

The predominance of entities such as 'HIV', 'WHO (World Health Organisation)', 'Nigeria', and various antiretroviral drugs underlines the international and local dimensions of the HIV/AIDS discourse in Nigeria. The presence of international entities like the World Health Organisation in the discourse emphasises the influence of international health policies and guidelines on regional practices and perceptions. The frequent mention of specific antiretroviral drugs points to a public knowledge base that could be leveraged to improve adherence to ART regimens and provide adequate access to these medications.

5.2 GEOGRAPHIC DISTRIBUTION INSIGHTS

The heat map analysis reveals critical insights into the geographical nuances of the HIV/AIDS discourse in Nigeria. The concentration of discussions in central and southwestern Nigeria, particularly around urban centres like Bauchi, Jos, Ibadan, and Enugu, suggests these areas as focal points for HIV/AIDS awareness and possibly higher rates of HIV prevalence or ART distribution challenges. The lesser activity in the northern areas could suggest a lower level of public discourse about HIV/AIDS, potentially due to cultural, social, or logistical reasons, or it may reflect a gap in data collection from these areas

5.2.1 STRATEGIC ESSENCES FOR ANTIRETROVIRAL DRUG DISTRIBUTION

The geographical insights the heat map provides are valuable for designing antiretroviral drug distribution schemes. Regions with high discussion intensity could be prioritised for interventions to harness the existing public engagement and address the high need or awareness. Conversely, areas with lower discussion intensity, particularly in the northern regions, may require targeted awareness campaigns and further investigation to understand the underlying reasons for the limited discourse and potentially unmet needs.

5.2.2 LIMITATIONS AND FUTURE DIRECTIONS

While the study leverages innovative Geo-NLP approaches to extract valuable insights from unstructured text data, there are inherent limitations. The reliance on social media and news data may not fully capture the standpoints and necessities of populations with limited internet access or those who do not participate in public discourse due to stigma or other barriers. Future research could integrate additional data sources, such as direct community engagement and health facility data, to provide a more comprehensive picture of Nigeria's antiretroviral drug distribution landscape. The study's findings also emphasise the potential for integrating Geo-NLP insights with other data streams, such as epidemiological and healthcare infrastructure data, to develop a multi-layered approach to health intervention planning and implementation.

6 CONCLUSION

This study introduced a groundbreaking method to enhance antiretroviral drug distribution in Nigeria by employing Geographic Natural Language Processing (Geo-NLP) to analyse social media and news data. A specialised Named-Entity Recognition (NER) model was crafted to sift through unstructured text, unveiling patterns in public discourse related to HIV/AIDS and antiretroviral therapy. Key findings included identifying areas with heightened discussions and pinpointing potential sites for focused health interventions. The study showcased the utility of Geo-NLP in shaping public health strategies and underscored the richness of social media and news as data reservoirs. Despite its innovations, the study also recognised its limitations, particularly the risk of overlooking less represented demographics. It highlighted the necessity of melding these insights with other data forms to paint a fuller picture of the health landscape, contributing to the global fight against the HIV/AIDS epidemic and propelling public health research forward through cross-disciplinary cooperation.

REFERENCES

Conway, M., Hu, M., & Chapman, W. W. (2019). Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and ConsumerGenerated Data. *Yearbook of Medical Informatics*, 28(01), 208–217. https://doi.org/10.1055/s-0039-1677918

Dirlikov, E., Jahun, I., Odafe, S. F., et al. (2021). Rapid Scale-up of an Antiretroviral Therapy Program Before and During the COVID-19 Pandemic — Nine States, Nigeria, March 31, 2019–September 30, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 70(12), 421–426. https://doi.org/10.15585/mmwr.mm7012a3

Federal Ministry of Health. (2020). Nigeria Essential Medicines List 2020 7th Edition.

Gammino, V. M., Nuhu, A., Chenoweth, P., et al. (2014). Using Geographic Information Systems to Track Polio Vaccination Team Performance: Pilot Project Report. *Journal of Infectious Diseases*, 210(suppl 1), S98–S101. https://doi.org/10.1093/infdis/jit285

National Agency for the Control of AIDS (NACA). (2021). NATIONAL HIV AND AIDS STRATEGIC FRAMEWORK 2021-2025. https://www.naca.gov.ng/wp-content/uploads/2022/03/National-HIV-and-AIDS-Strategic-Framework-2021-2025-Final.pdf

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labelling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3), 671–681. https://doi.org/10.1093/jamia/ocu041

Olutuase, V. O., Iwu-Jaja, C. J., Akuoko, C. P., Adewuyi, E. O., & Khanal, V. (2022). Medicines and vaccines supply chains challenges in Nigeria: a scoping review. *BMC Public Health*, 22(1), 11. https://doi.org/10.1186/s12889-021-12361-9