GEO-SEMANTIC ANALYSIS OF MEDICAL RESEARCH TRENDS IN NIGERIA

Ahmad Ibrahim Ismail, Anthony Soronnadi, Olubayo Adekanbi, Bashirudeen Ibrahim & David Akanji
Data Science Nigeria (DSN),
Lagos, Nigeria.
{ahmad,anthony,olubayo,bashirudeen,david}@datasciencenigeria.ai

ABSTRACT

In the context of a rapidly evolving global health landscape, this study aims to cast light on the focal points and regional intricacies of medical research in Nigeria. It addresses the critical need to align medical research with health policies, responding to the dynamic health requirements of Nigeria's diverse population. Utilizing a Geo-semantic approach, the research melds Geospatial Analysis with the advanced capabilities of Natural Language Processing. This methodology was applied to analyze and visually interpret Nigerian medical research's thematic and geographic trends based on articles from the PubMed database. The study uncovered distinct regional focuses and collaborative networks in medical research, underscoring the importance of aligning research efforts with the prevalent health challenges. The study found emergent challenges like COVID-19 and epidemiological studies receiving optimum attention, while prevalent health challenges like health insurance and neglected tropical diseases were on the dwindling end of research interest. These findings provide a blueprint for improving the effectiveness of medical research and healthcare policy in Nigeria, offering significant insights for strategic planning and resource allocation in the health sector. Moreover, this innovative approach demonstrates the feasibility and value of integrating NLP and geospatial analysis in medical research. It opens new avenues for low- and middle-income countries to derive insights and enhance their healthcare planning strategies by leveraging data from unstructured sources.

1 INTRODUCTION

The increased need for health solutions, coupled with technological advancements, has propelled an expansion of medical research in modern times, leading to the creation of a vast repository of scientific knowledge (Harrison & Sidey-Gibbons, 2021). This has presented opportunities for researchers and innovators to leverage on these repositories for diverse use cases, and also some challenges for researchers looking to understand this rapidly evolving landscape, e.g., lack of representation for some locally published works. This study embarks on an innovative exploration of the focus of research in the Nigerian medical landscape, where unique socio-economic and health challenges often shape the dynamics of medical research.

The study brings together a unique blend of traditional web scraping, Geo-semantics, and the growing field of Natural Language Processing (NLP) to delve into the massive database of PubMed Central to extract, analyze, and understand the geographical distribution and thematic concentration of medical research in Nigeria. Geospatial semantics (Geo-semantics) is a field that focuses on the representation of geographic entities in both cognitive and digital utilizations. It aims to enhance the design and interoperability of geographic information systems (GIS) and aid user interaction in handling big geo-data and unstructured text via advanced methods like natural language processing (Hu, 2018). Converting unstructured data (e.g., research articles) into structured geospatial information involves toponym recognition using Named Entity Recognition (NER) and toponym resolution using geocoding tools. This process outputs the geospatial identification (longitudes and latitudes) of extracted locations, which this study utilized to explore the focus of medical research in Nigeria.

2 RELATED WORKS

Many research works have used different NLP techniques and Geospatial analysis to assess and extract insight from openly available unstructured data. A bibliometric analysis of NLP in medical research reported an 18.3 percent annual growth rate for using NLP techniques in medical research between 2007 and 2016 (Chen et al., 2018). This underscores the significant possibilities NLP methods and applications offer for processing medical information. A study used NLP techniques to conduct a lexicon-based sentiment analysis of selected drugs using data collected from pharmaceutical review websites (Harrison & Sidey-Gibbons, 2021). While the research does not utilize geospatial information, it showcased a comprehensive analysis of unstructured medical data using NLP. Researchers leveraged Bidirectional Encoder Representations from Transformers (BERT), a prominent NLP model, and qualitative content analysis together with geospatial time series modeling in the United States (US) to analyze public sentiments on social media regarding COVID-19 vaccination and their geographic pattern (Ye et al., 2023). The study reported how opinions regarding COVID-19 vaccination changed over time in different locations and demonstrated the potential of an analysis pipeline incorporating NLP-enabled models, time series, and geographical analysis to enable real-time, large-scale, and trustworthy analyses that could address public health concerns and provide foundations for data-driven health policies and communication strategies.

This work aims to build on these to provide a broad geographic overview of medical research in Nigeria using open-access research articles to provide a novel perspective on the current state and future directions of medical research in Nigeria. It aims to enable a more targeted and effective healthcare strategy and policy making in a diverse nation like Nigeria.

3 RESEARCH METHODOLOGY

The study combined natural language processing (NLP) and geospatial exploratory data analysis to investigate the medical research focus from original research articles with open access tags on PubMed Central.

3.1 DATA ACQUISITION

PubMed Central is a free, full-text biomedical and life sciences journal literature archive. It is one of the largest digital research repositories in the world, maintained by the National Center for Biotechnology Information (NCBI). The website's advanced search function was used to find articles mentioning "Nigeria" and filtered for January 2022 to November 2023. This period was selected due to limited resources and time, as the number of files to download increases exponentially with each year. The filter results were saved in the PubMed format, which could be queried for all information about an article except the full text. The 'biopython' module was used to query the downloaded file to extract articles' PMIDs, PMCID, title and abstract which were saved into a data frame using the 'pandas' module. The PMCID was then used to scrape each article in PDF format. Python's 'PyPDF2' library was used to extract the full text from all the downloaded article PDFs into separate text documents identified by their PMCID.

3.2 DATA PREPARATION

To extract the research location, the research methodology was queried for mention of locations in the methodology section of each paper. This method efficiently filtered out other types of research articles like commentaries, letters to Editors, Conference Reports etc. (which wouldn't typically contain a "Methodology" section), while also excluding locations mentioned in the introductions and literature reviews.

Named Entity Recognition and Information Retrieval: "WikiNEuRal"[1], a novel NER model that exploited multilingual BERT's architecture to distinguish named entities from concepts, val-

[1] https://huggingface.co/Babelscape/wikineural-multilingual-ner

idate annotations, and discover annotations, was used to extract the locations from the research methodologies (Tedeschi et al., 2021).

The extracted locations were added to the dataset containing information on the articles. Labels were manually created and assigned by the research team to capture the medical areas or focuses of the research work to obtain the research focuses. These labels are explained in Table 1 below:

Table 1: Definition of Research Categories.

s/n	Label	Explanation
1	Behavioral Health	All research that relates to understanding the beliefs, attitudes, and perceptions of a target population and how they affect the specific research focus (e.g., diseases or programs). Examples include: "Dental Health Knowledge, Attitude, and Practice Among University of Calabar Students."
2	Cancer	Research relating to cancer biology, prevention, diagnosis, and treatment.
3	Chronic Diseases	Research focusing on long-term health conditions alone without specific combinations with infectious diseases, therapy, or epidemiology. e.g., "Oral innate immunity in patients with type 2 diabetes mellitus in a tertiary hospital in Ibadan Nigeria: a cross-sectional study.".
4	COVID-19	All research relating to the COVID-19 pandemic and its impact.
5	Drug Research and Development	Research focused on drug design, development, and evaluation including the therapeutic potential of plants. E.g., "Evaluation of the haematinic, antioxidant and anti-atherosclerotic potential of Momordica charantia in cholesterol-fed experimental rats."
6	Environmental Health and Safety	Research relating to environmental variables and their effects on human health, including research relating to safety provisions in communities.
7	Epidemiology	All research relating to the distribution and determinants of diseases and health-related events (excluding COVID-19 and HIV/AIDS).
8	Health Insurance	Research on the Nigerian National Health Insurance Scheme (NHIS) and universal health coverage.
9	Health Policy	Research relating to the impact of policies on the health of communities and individuals.
10	HIV/AIDS	All research relating to HIV/AIDS including its impact, management, and concomitance with other diseases.
11	Medical Imaging	Research relating to medical imaging technologies, techniques, and applications.
12	Medical Training	Research relating to medical training, education, and professional development. E.g., "Assessment of the Adequacy of Neurosurgery Teaching Methods among Medical Students in Enugu State, Nigeria."
13	Neglected Tropical Diseases (NTDs)	Research focusing on the NTDs, including their epidemiology, prevention, treatment, and control strategies.
14	Nutrition, Dietetics, and Disease Prevention	Research focusing on nutrition, nutrition deficiency, and prevention of diseases using nutrition (separate from child nutrition).
15	Occupational Health	Research focusing on workplace hazards, occupational diseases, safety measures, and the impact of work on health.
16	RMNCHN	All research around Reproductive, Maternal, Newborn, and Child Health and Nutrition.
17	Therapeutic Research	All research relating to the development, evaluation, improvement, and safety of therapeutic interventions including medications, medical treatments, and healthcare approaches.

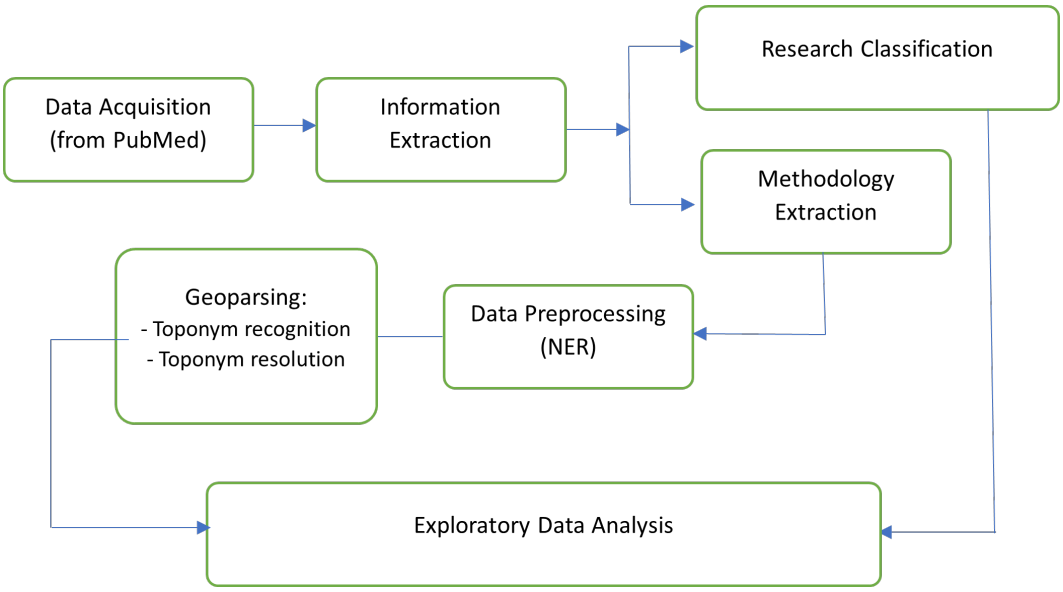

Figure 1: Research Flowchart

Geoparsing: This entails two steps:

1. Toponym Recognition: After extracting the locations using the NER model, a dataset containing the states, capitals, and major cities in the country was also used to filter them. This step was necessary to ensure that the focus on 'Nigeria' was maintained, and all unrelated locations removed.

2. Toponym Resolution: The Python 'geocoder' module was used to geocode identified locations. This was able to resolve almost all of the local towns and city locations in the dataset. The others were done manually with more location contexts using google maps website.

3.3 Exploratory Data Analysis

The final dataset used for analysis contained the location mentioned in the research methodologies, their geospatial data (longitude and latitude), and the research focus (labels). Popular python libraries, geopandas and matplotlib, were utilized to conduct the exploratory data analysis (EDA), and folium was used for geospatial visualizations.

4 Results and Discussion

The exploratory data analysis (EDA) of filtered medical research articles from PubMed Central focusing on the Nigerian original research articles revealed patterns in research focus and geographical distribution of research. The dataset had 1502 entries from 1000 research papers detailing the research focus, location, and corresponding geographical coordinates.

4.1 Research Focus Distribution

Analysis of the research focus revealed that a diverse range of medical research topics were being investigated in Nigeria. The 17 unique labels used to identify the research focus areas had varying degrees of research activity. The distribution of these focuses provides a clear indication of the prevailing medical research interests in the country, highlighting areas of both high and low concentrations. Epidemiological research is the most common, followed closely by RMNCHN research; these studies are typical of low- and middle-income countries; this is shown in the figure 2 below:

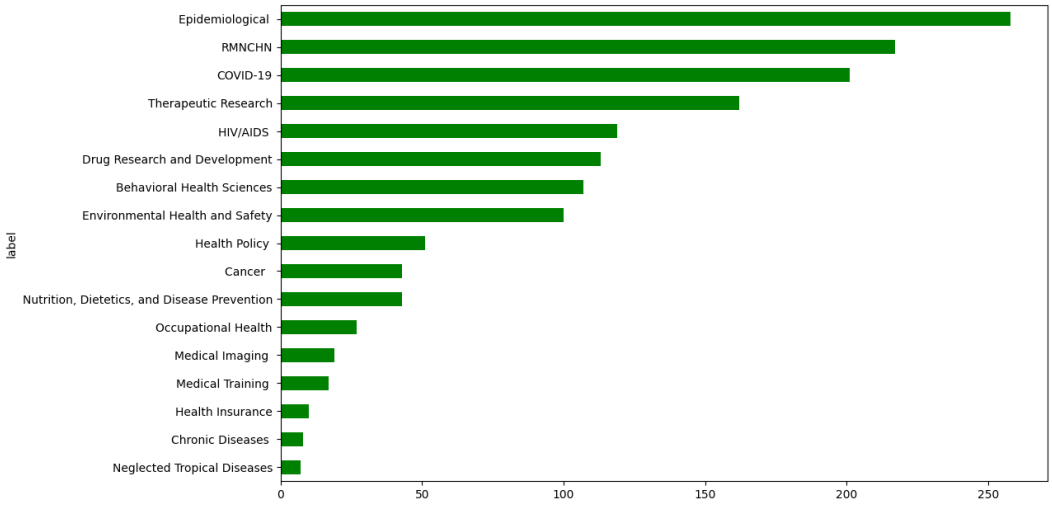

Figure 2: Research Focus Distribution

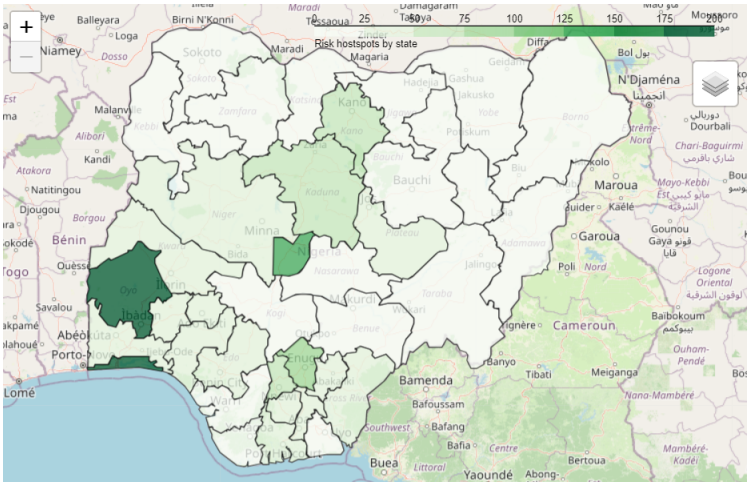

Figure 3: Map of Research Distribution

4.2 Geographical Distribution of Research

The analysis revealed the spatial distribution of original research in Nigeria between January 2022 and November 2023. Lagos state and Oyo state stand out as the country's major loci of research. These are followed by the Federal Capital Territory, Enugu, Anambra, Kano, and Kaduna states. This may not come as surprising as they're among the most populous states in the country and are home to some of the most prominent medical research institutions in the country like Lagos State University and University of Ibadan (and the teaching hospitals). The uneven distribution also suggests an alignment of research activities with specific regional health challenges and the presence of established research institutions. Figure 3 below shows the distribution of research articles in the studies by state.

The analysis highlighted a diverse range of medical research topics with varying degrees of concentration, reflecting the country's current health priorities and research interests. The research areas with lesser focus, like medical imaging, medical training, chronic diseases, and NTDs, present opportunities for exploration. Chronic diseases were found to be majorly studied as concomitant illnesses, leading to them being categorized under other labels, e.g., "Clinical characteristics and treatment patterns of pregnant women with hypertension in primary care in the Federal Capital

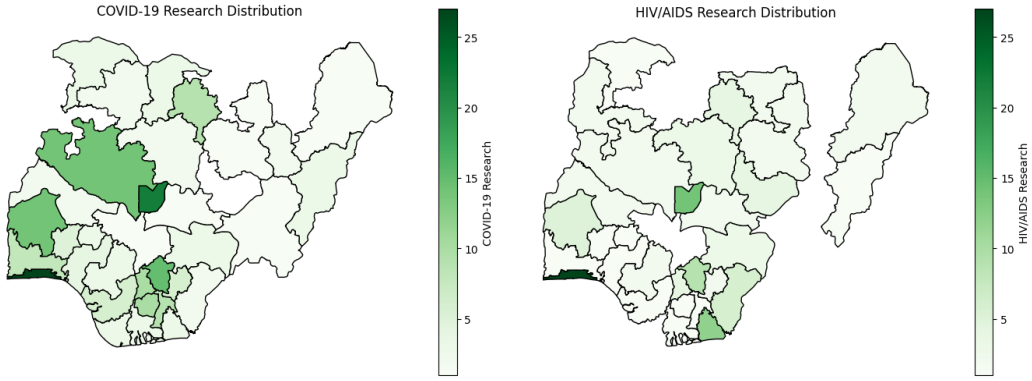

Figure 4: COVID-19 and HIV/AIDS Research Maps

Territory of Nigeria: cross-sectional results from the hypertension treatment in Nigeria Program" was categorized under RMNCHN, as the study population was pregnant women; and "Factors influencing the integration of evidence-based task-strengthening strategies for hypertension control within HIV clinics in Nigeria" was under the HIV/AIDS label. Another critical factor that could contribute to a seemingly lesser focus on some research areas, like medical imaging, occupational health, and medical training research, is that some of them are published in local journals that are not "selected as a MEDLINE journal or deposited to the PMC" (Huh, 2016). This lack of inclusion in prominent global databases can result in underrepresentation in international literature reviews and analyses. This phenomenon is not unique to Nigeria; it is a challenge faced by many countries where local journals, while potentially of high quality and relevance, do not meet the criteria (like impact factor, peer review standard, regularity of publication, etc.) for inclusion in major international databases. As a result, valuable research conducted in specific fields and regions remains less visible on the global stage (Ofori-Adjei et al., 2006).

The decision not to focus on the publishing date of research articles in the 23-month timeframe (January 2022 to November 2023) of this study considers the intricacies involved in the scientific publication process. This process is often lengthy and complex, encompassing various stages from the initial theoretical conception of research to its final preparation for publication. Challenges such as journal rejections and revisions can significantly extend the timeline for research to appear in scientific journals. Therefore, emphasizing the publication date could misrepresent the actual time and effort invested in the research development. This consideration is crucial for accurately understanding and interpreting trends and focuses in the selected body of research.

Some important research focus are discussed below:

4.3 COVID-19 AND HIV/AIDS

Due to the nature of these diseases, they were each allotted a category of their own. All research focusing on the epidemiology of these diseases and their impact, including studies focusing on them as concomitant illnesses, were categorized under their labels. This led to them standing out as the third and fifth most occurring research focus in this study. The COVID-19 global pandemic has profoundly impacted public health, research, and policies worldwide, and Nigeria was no exception. The studies predominantly concentrated on epidemiology, impact assessment, and management strategies. The geographical spread of COVID-19 research was extensive, reflecting the nationwide impact of the pandemic. There is a high concentration of COVID-19 research in urban centers like Lagos and Abuja, likely due to the higher population density, which led to a more significant number of cases and, consequently, a greater need for research in these areas. Several studies focused on the resilience and responsiveness of the Nigerian healthcare system as well as the socio-economic impact of the pandemic.

HIV/AIDS is a longstanding public health challenge with extensive research coverage. The distribution of research on HIV/AIDS varied across different regions, with studies focusing on prevention

and treatment as well as co-morbidities and concomitant illnesses. The research on HIV/AIDS in Nigeria exhibited regional variability, with an intensified focus on areas like Akwa Ibom and the FCT, known for their higher HIV prevalence rates. This geographical concentration of studies aligns with prevention strategies, antiretroviral treatment regimes, and community health education initiatives (Awofala & Ogundele, 2018). The exploration of HIV as a co-morbidity factor, particularly in the context of prevalent diseases such as tuberculosis and some chronic diseases, underscored the multifaceted health challenges faced by those living with HIV/AIDS, pointing to the need for a comprehensive approach in both treatment and prevention.

4.4 Health Insurance

Health insurance is essential to achieving universal health coverage. In Nigeria, the National Health Insurance Scheme (NHIS) was tasked with this mandate at its establishment in 1999. Despite this and many more recent efforts, health insurance and, by extension, universal health coverage in Nigeria is still a dream for Nigerians, where most of the 95 percent of the population not covered by the NHIS rely on out-of-pocket payment for their healthcare needs (Oxford Business Group, 2023). Several obstacles have hindered the desired progress, including cultural attitudes, poor health service delivery, distrust in public services, exclusion of primary healthcare centers from NHIS coverage, and programmatical exclusion of rural and remote populations. These have led to poor uptake of health insurance, most notably among rural dwellers and workers in the informal sector (Mariam, 2022). The private health insurance industry, which caters to a limited portion of the population, has continued to grow. This is not to say that the private sector doesn't also grapple with its own challenges, including high insurance premiums, medical inflation, and complex health policies (Mariam, 2022). With the significant gap in understanding and addressing the complexities of health insurance in Nigeria, the under-researched status of health insurance in Nigeria identified by this research emphasizes the need for more focused studies to support policymaking and implementation and achieve universal health coverage in Nigeria.

4.5 Neglected Tropical Diseases

NTDs pose significant health risks to underprivileged populations in Nigeria. Although the country has recorded historical progress thanks to research and collaborations by the Nigerian Institute for Medical Research (NIMR), international organizations continue to call for more stakeholder attention to the challenges preventing the country from meeting the World Health Organization (WHO) recommendations for some ailments like schistosomiasis, trachoma, and soil-transmitted Helminthes (The Conversation — Gavi VaccinesWork, 2023). These results also highlighted the need for more focused research that could significantly aid in understanding and overcoming the challenges faced in achieving the NTD recommendations by WHO. This could also provide insights into more effective strategies for control and elimination, address barriers, and raise awareness about these diseases.

5 Conclusion

This research conducted an analysis of Nigerian medical research between January 2022 and November 2023. The study revealed a landscape rich in diversity but marked by disparities. Key areas COVID-19, HIV/AIDS, and RMNCHN received significant attention, reflecting the global and national priorities. Epidemiological studies were the most popular, which is quite typical of Nigeria as a developing country in the tropical region of the globe. Under-represented areas, including health insurance and NTDs, are at the far end of the spectrum. These are particularly relevant considering the complexities of the health insurance landscape in Nigeria and the ongoing struggles against NTDs. The challenges in these areas, such as low coverage, unfavorable cultural beliefs, low government health spending, etc., emphasize the need for more focused research. This research thus underscores the importance of aligning medical research with both current and emerging health challenges, with researchers needing to balance response to immediate health crises like COVID-19 with necessary attention to systemic issues like health insurance and NTDs. This research also acknowledges the need for more works published in local journals not featured in PubMed to be added to international repositories. Seemingly under-researched areas like medical training and medical

imaging research may be victims of this challenge, thus affecting their placement in this and potentially similar work, looking to use these international repositories, which are openly accessible.

REFERENCES

Awoyemi Abayomi Awofala and Olusegun Emmanuel Ogundele. HIV epidemiology in Nigeria. *Saudi Journal of Biological Sciences*, 25(4):697–703, may 2018. ISSN 1319562X. doi: 10.1016/j.sjbs.2016.03.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1319562X16300110.

Xieling Chen, Haoran Xie, Fu Lee Wang, Ziqing Liu, Juan Xu, and Tianyong Hao. A bibliometric analysis of natural language processing in medical research. *BMC Medical Informatics and Decision Making*, 18(S1):14, mar 2018. ISSN 1472-6947. doi: 10.1186/s12911-018-0594-x. URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0594-x.

Conrad J. Harrison and Chris J. Sidey-Gibbons. Machine learning in medicine: a practical introduction to natural language processing. *BMC Medical Research Methodology*, 21(1):158, dec 2021. ISSN 1471-2288. doi: 10.1186/s12874-021-01347-1. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01347-1.

Yingjie Hu. Geospatial Semantics. *Comprehensive Geographic Information Systems*, (2017):80–94, 2018. doi: 10.1016/b978-0-12-409548-9.09597-x.

Sun Huh. How to Add a Journal to the International Databases, Science Citation Index Expanded and MEDLINE. *Archives of Plastic Surgery*, 43(06):487–490, nov 2016. ISSN 2234-6163. doi: 10.5999/aps.2016.43.6.487. URL http://www.thieme-connect.de/DOI/DOI?10.5999/aps.2016.43.6.487.

Adetona Mariam. Two decades later, Nigeria's health insurance is still flailing, may 2022. URL https://www.aljazeera.com/features/2022/5/11/two-decades-after-nigerias-health-insurance-is-still-flailing.

David Ofori-Adjei, Gerd Antes, Prathap Tharyan, Elizabeth Slade, and Pritpal S Tamber. Have Online International Medical Journals Made Local Journals Obsolete? *PLoS Medicine*, 3(8):e359, aug 2006. ISSN 1549-1676. doi: 10.1371/journal.pmed.0030359. URL https://dx.plos.org/10.1371/journal.pmed.0030359.

Oxford Business Group. Mandatory health insurance to ease access to care in Nigeria. Technical report, 2023. URL https://oxfordbusinessgroup.com/reports/nigeria/2022-report/economy/eye-on-the-prize-legislation-making-health-insurance-mandatory-is-the-latest-step-

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2521–2533, Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.215. URL https://aclanthology.org/2021.findings-emnlp.215.

The Conversation — Gavi VaccinesWork. 100 million Nigerians are at risk of neglected tropical diseases: what the country is doing about it. Technical report, 2023. URL https://www.gavi.org/vaccineswork/100-million-nigerians-are-risk-neglected-tropical-diseases-what-country-doing-abou

Jiancheng Ye, Jiarui Hai, Zidan Wang, Chumei Wei, and Jiacheng Song. Leveraging natural language processing and geospatial time series model to analyze COVID-19 vaccination sentiment dynamics on Tweets. *JAMIA Open*, 6(2), apr 2023. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooad023. URL https://academic.oup.com/jamiaopen/article/doi/10.1093/jamiaopen/ooad023/7116330.