https://doi.org/10.51408/issi2025_176

Machine learning-based model to predict topics contributing to Sustainable Development Goals: A study of Latin American and European Countries

Barbara S. Lancho Barrantes

b.lanchobarrantes@brighton.ac.uk
School of Architecture, Technology and Engineering, University of Brighton
Brighton BN2 4GJ (United Kingdom)

Introduction

Health is a human right and a cornerstone of physical, mental, and social well-being (WHO, 1949). Ensuring access to healthcare is not only ethical; it is essential for reducing poverty and fostering inclusive, sustainable development.

Yet, health systems worldwide face persistent challenges: underfunding, high out-of-pocket costs, fragmented service delivery, and inequities based on the ability to pay. These issues, compounded by ineffective governance, undermine progress toward universal health coverage and financial protection.

The 2030 Agenda for Sustainable Development, adopted by the United Nations in 2015, includes 17 SDGs, with Goal 3 dedicated to ensuring healthy lives and well-being for all. Though interlinked with other goals, SDG 3 plays a pivotal role in shaping global health priorities.

Recent bibliometric studies have explored how countries' research aligns with SDG challenges (Yamaguchi et al. 2023)However, much of this work focuses on sectors like business or education, leaving a gap in understanding SDG 3-related research, especially in Global South and Global North countries (Yaqub, et al., 2024) Notably, low-income countries, despite facing the greatest SDG-related challenges, contribute minimally to the research driving global progress (Confraria et al., 2024).

This study uses the OpenAlex (Priem et al.2022) database to examine how countries' SDG 3 research priorities differ. It highlights the need for context-sensitive bibliometric strategies that account for regional health challenges often overlooked in global analyses.

By analysing large-scale publication data, research trends, topics, and collaboration patterns, AI tools can predict emerging health research themes with growing accuracy. The research questions of this study are:

- How do research priorities differ between Latin America (Global South) and Europe (Global North)?
- Do both regions focus on similar health challenges with equal intensity?
- Can machine learning-driven models effectively forecast future research trends?

Data and Methods

This study draws on data from OpenAlex, a large open-access bibliographic database with over 240 million scholarly works. The study focused on SDG 3: Good Health and Well-being, selecting the top 10 publishing countries from:

- Latin America: Brazil, Uruguay, Mexico, Colombia, Chile, Costa Rica, Puerto Rico, Argentina, Ecuador, and Peru.
- Europe: United Kingdom, France, Germany, Spain, Netherlands, Switzerland, Italy, Belgium, Sweden, and Poland.

OpenAlex leverages a machine learning-based SDG Classifier to assess the relevance of academic publications to the 17 Sustainable Development Goals (SDGs). Using Natural Language Processing (NLP), the system analyses titles, abstracts, keywords, and citations to understand the content of each

publication. The SDG BERT model, a multilingual, multi-label transformer trained on SDG-labeled data (Aurora Query Model v5), then assigns a probability score (ranging from 0 to 1) for each SDG, indicating the publication's relevance.

In addition to SDG tagging, OpenAlex employs an automated topic classification system that assigns each publication to one or more of ~4,500 scientific topics.

Results

Table 1 compares health research priorities and contributions to SDG 3 between Latin American and European countries, highlighting differences in focus and scientific output.

Publications count	Classification		
Topic	Europe	Latin America	Total
SARS-CoV-2 and COVID-19 Research	4954	953	5907
Cancer Immunotherapy and Biomarkers	3931	300	4231
COVID-19 and Mental Health	2678	1175	3853
Liver Disease Diagnosis and Treatment	3265	581	3846
Cardiac Valve Diseases and Treatments	3435	231	3666
COVID-19 Clinical Research Studies	2510	1038	3548
Inflammatory Bowel Disease	2981	220	3201
Atrial Fibrillation Management and Outcomes	2844	200	3044
Diabetes Treatment and Management	2655	326	2981
CAR-T cell therapy research	2919	9	2928
Long-Term Effects of COVID-19	2147	765	2912
Mosquito-borne diseases and control	1291	1353	2644
Pancreatic and Hepatic Oncology Research	2402	239	2641
Prostate Cancer Treatment and Research	2348	196	2544
COVID-19 and healthcare impacts	2020	491	2511
Lung Cancer Treatments and Mutations	2306	137	2443
Acute Myeloid Leukemia Research	2266	177	2443
Tuberculosis Research and Epidemiology	1848	592	2440
Glioma Diagnosis and Treatment	2132	232	2364
Acute Ischemic Stroke Management	1968	362	2330

Europe leads in publications on advanced medical topics like cancer and cardiac diseases, while Latin America focuses on mosquito-borne diseases, mental health, and long-term pandemic effects. COVID-19 research is a shared focus, with Europe concentrating on clinical studies and Latin America on mental and social health impacts. These differences reflect regional health priorities and socioeconomic factors.

A Random Forest classifier is used to predict publication origin based on research topics. The model's accuracy improved with more data but was affected by class imbalance, as Europe had more publications, potentially introducing bias.

The model shows a moderate ability to distinguish between regional research focuses, but performance is limited. Despite a balanced dataset (Europe: 422, Latin America: 378), topic overlap likely reduced prediction clarity. Results suggest that thematic differences exist but are not strongly distinctive based on topic data alone.

```
Model: Logistic Regression
Accuracy: 0.55625
Classification Report:
              precision    recall  f1-score   support

      Europe       0.59      0.53      0.56       422
Latin America      0.53      0.58      0.55       378

    accuracy                           0.56       800
   macro avg       0.56      0.56      0.56       800
weighted avg       0.56      0.56      0.56       800

------------------------------------------------------------
Model: Random Forest
Accuracy: 0.575
Classification Report:
              precision    recall  f1-score   support

      Europe       0.60      0.58      0.59       422
Latin America      0.55      0.57      0.56       378

    accuracy                           0.57       800
   macro avg       0.57      0.57      0.57       800
weighted avg       0.58      0.57      0.58       800

------------------------------------------------------------
Model: Naive Bayes
Accuracy: 0.55125
Classification Report:
              precision    recall  f1-score   support

      Europe       0.59      0.48      0.53       422
Latin America      0.52      0.63      0.57       378

    accuracy                           0.55       800
   macro avg       0.56      0.56      0.55       800
weighted avg       0.56      0.55      0.55       800

------------------------------------------------------------
```

Future research should focus on individual countries' contributions to SDG 3 rather than on regions. Analysing publication abstracts and exploring models like Gradient Boosting (e.g., XGBoost, LightGBM) and deep learning-based NLP could reveal subtle patterns. Additionally, optimising the Random Forest model with techniques like Grid or Random Search could enhance performance.

Conclusions

The number of publications on SDG 3: Good Health and Well-Being has increased significantly in the last two years, creating challenges for researchers to identify priorities among countries. This study used machine learning and NLP to track shifts in health research topics, from "SARS-CoV-2" to emerging areas like "Zika" and "Cancer." It

highlights gaps in research on topics like "Palliative Care" and "Cerebral Venous Sinus Thrombosis." By analysing open research data, the study predicted future trends for 10 Latin American and 10 European countries, revealing ongoing regional differences in health priorities. It emphasises the need for stronger partnerships, more funding, and improved capacity in Latin America. Overall, machine learning and NLP enhance research efficiency and support decision-making for SDG 3.

References

Confraria, H, Ciarli, T & Noyons, E, (2024). Countries' research priorities in relation to the Sustainable Development Goals. Research Policy, Elsevier, vol. 53(3).

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833

Yaqub, O, Coburn, J & Moore, D A.Q. (2024). Research-targeting, spillovers, and the direction of science: Evidence from HIV research-funding. Research Policy, 53(8).

Yamaguchi, N. U., Bernardino, E. G., Ferreira, M. E. C., de Lima, B. P., Pascotini, M. R., & Yamaguchi, M. U. (2023). Sustainable development goals: A bibliometric analysis of literature reviews. Environmental Science and Pollution Research, 30(3), 5502–5515.