Research Output vs Country Development Indicators (2023)

AAAA, IT University of Copenhagen, Denmark

ACM Reference Format:
AAAA. 2018. Research Output vs Country Development Indicators (2023). *ACM/IMS J. Data Sci.* 37, 4, Article 111 (August 2018), 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Scientific research is the primary engine of global innovation, yet its production is geographically uneven. While it is well-established that large economies like the United States and China dominate in terms of total publication volume, the drivers of research efficiency are less understood. This project investigates the relationship between a nation's development and its research output across three distinct fields: Artificial Intelligence (AI), Environmental Science, and Medicine.

The core motivation for this analysis is to move beyond simple rankings of scale. By focusing solely on the total number of publications, we often obscure the achievements of smaller nations that may be far more effective at mobilizing their limited resources. This report seeks to decouple scientific success from simple population size and economic mass, aiming instead to build a profile of research intensity. We explore whether the ingredients for success are universal, meaning they are simply a result of national wealth, or if different scientific domains require different national characteristics, such as specific digital infrastructure for AI or public health outcomes for Medicine.

To achieve this, we integrated real-world data from multiple open sources to construct a comprehensive view of the global research landscape in 2023. Rather than analyzing these metrics in isolation, we examined how a country's socioeconomic backdrop correlates with its scientific output. This approach allows us to distinguish between the "Giants" of research, who dominate through sheer magnitude, and the "Specialists" who achieve high efficiency through focused development.

2 Data Collection

To build the dataset for this analysis, we combined information from two open sources. We needed one source for the "input" variables,the economic and social factors, and another for the "output" variable—the scientific research itself.

For the socioeconomic data, we used the World Bank Open Data platform. We selected eight indicators that provide a broad picture of a country's status, including economic measures like GDP and GDP per capita, but also social metrics like life expectancy, internet access, and tertiary school enrollment. The data was downloaded as a raw CSV file, which required some initial filtering to separate actual countries from the regional aggregate entries often found in World Bank datasets.

Author's Contact Information: AAAA, IT University of Copenhagen, Copenhagen, Denmark, aaaa@asas.com.

ACM 2831-3194/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

For the research data, we used OpenAlex, an open index of the global research system. Unlike the static World Bank files, this data had to be retrieved programmatically. We wrote a Python script to query their API, specifically looking for works published in 2023. We retrieved the total publication count for every country. Then, we filtered these queries by specific fields to isolate counts for Artificial Intelligence, Environmental Science, and Medicine.

3 Data Processing and EDA

Working with "wild" data presents unique challenges that require rigorous cleaning and transformation before any analysis can begin. Our raw datasets, particularly from the World Bank, were not ready for immediate integration. We implemented a multi-stage processing pipeline to ensure data quality and consistency.

(1) Cleaning Aggregates
The primary challenge with the World Bank dataset was that it mixed individual countries with broad regional aggregates. Upon initial inspection, we identified 31 rows that did not represent sovereign nations but rather groupings such as "Arab World," "High income," and "Sub-Saharan Africa". Including these would have severely skewed our analysis, essentially double-counting populations and GDP.
We systematically filtered these out by identifying their unique ISO-3 codes (e.g., ARB for Arab World, HIC for High Income) and removing them from the dataset. This critical step reduced our dataset from a raw dump to a clean list of 217 distinct

(2) Geographic Standardization
To enable regional analysis, we needed to map every country to its respective continent, so we constructed a dictionary mapping ISO-3 codes to their continental regions (Africa, Asia, Europe, North America, Oceania, South America).
By creating a continent mapping for all countries, we achieved we were able to later color-code our visualizations by continent and identify regional trends in research output.
That is a perfect addition. It bridges the gap between "we cleaned the data" and "we used MICE" by showing why MICE was a valid choice (because the variables are actually correlated).
Here is the text for that specific part of Section 3: Data Processing. You can insert this right before you talk about the MICE imputation code.

(3) Feature Correlation and Imputation Strategy
Before addressing the missing values in our dataset, we performed a preliminary correlation analysis to understand the relationships between our development indicators. We hypothesized that a country's infrastructure metrics (like internet access) would be strongly predictive of its social metrics (like education and health), which would allow us to accurately estimate missing data points.(1)

To ensure the integrity of our analysis, we adopted a systematic, column-by-column inspection strategy. Rather than applying broad automated filters, we checked each development indicator individually to identify data irregularities.

It is often discouraged relying solely on univariate inspection because it fails to detect multivariate outliers—data points that appear normal in isolation but are anomalous when combined. Usually, multivariate outlier detection is preferred to identify and remove these complex errors. However, in our context of global geopolitical data, "outliers" are almost never measurement errors—they are sovereign nations with distinct, extreme characteristics. For example, the United States (high GDP) and China (high Population) are statistical outliers in almost every dimension. A multivariate detection algorithm might flag these unique profiles as "anomalies" to be removed. Removing these data points would exclude the very "Giants" of research we aim to study, rendering our analysis of global trends incomplete. Therefore, we consciously rejected multivariate exclusion in favor of univariate transformation and imputation, prioritizing the preservation of every country in our dataset.

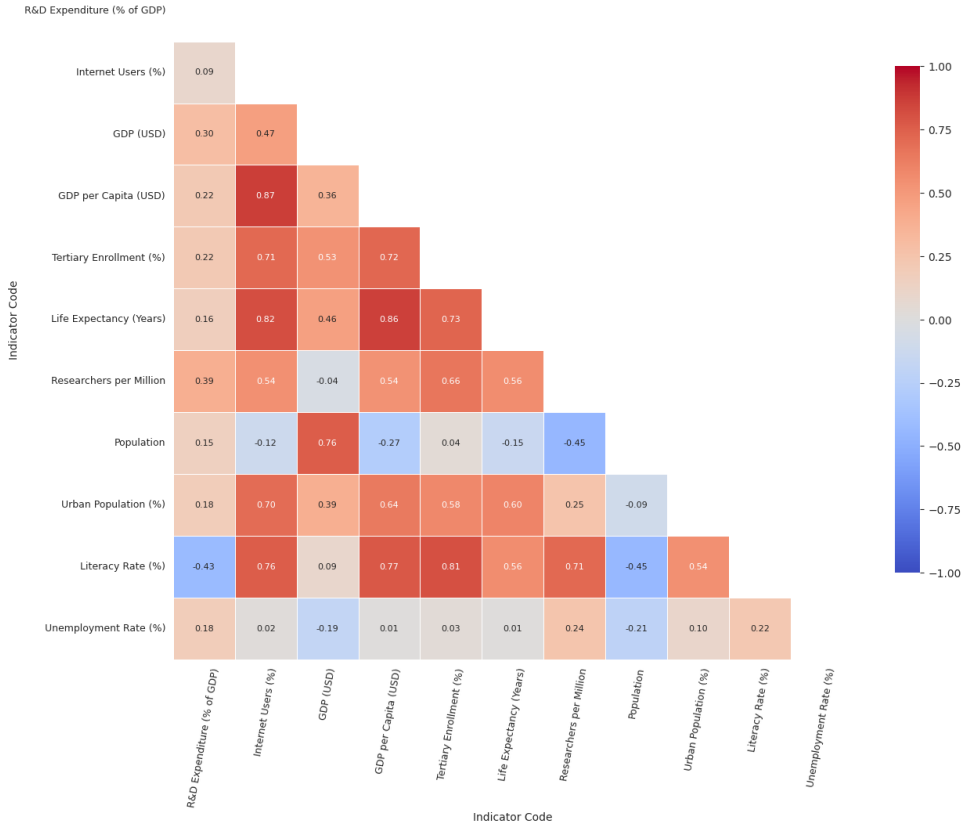

Fig. 1. Heatmap Correlation between the development indicators

1. Handling Distributional Skew (Transformation): The column-by-column inspection highlighted extreme skew in our "volume" metrics (GDP, Population), where a few nations dominated the scale. Rather than treating these extreme values as noise to be deleted, we treated them as valid signals that needed scaling. We applied a logarithmic transformation (np.log1p) to these specific columns. This approach "pulled" the extreme outliers back into a comparable range without losing the valuable information they represent, normalizing the distributions for our correlation analysis while preserving the rank order of the nations.

2. Handling Missing Values : Consistent with our philosophy of data preservation, we rejected the standard approach of deleting rows with missing data. Doing so would have introduced significant bias, likely excluding developing nations from our analysis.

Instead, we chose to go with a hybrid imputation strategy tailored to the nature of each variable. We prioritized context-aware methods over simple global means. For indicators where our initial analysis revealed strong correlations, we used those highly correlated features to define similarity groups (such as wealth bins) and imputed values based on those group medians. For metrics driven by location, we utilized regional imputation, filling gaps using continental medians. Finally, where possible, we avoided estimation entirely by calculating

values directly from their available underlying components (e.g. deriving per-capita metrics from total population data).

The most significant challenge was Tertiary Enrollment (%), which had the highest proportion of missing values. Before imputing, we investigated the nature of this missingness to determine if it was systematic. By plotting the distributions of other key variables (like Life Expectancy and GDP) for countries with and without enrollment data, we observed clear patterns, identifying the data as Missing At Random (MAR).

This confirmation allowed us to employ Multivariate Imputation by Chained Equations (MICE). Because the missingness was related to observed data, MICE could effectively model the missing values as a function of the other highly correlated features (Life Expectancy, Internet, GDP). We ran the algorithm for 10 iterations, ensuring that the imputed education data was statistically consistent with the country's entire socioeconomic profile.

Finally, for indicators where the missingness was too extensive to be reliably imputed,(Research and Development Expenditure,Literacy Rate (%),Researchers per Million) which was missing for the majority of nations—we made the decision to drop the column entirely.

4 Methods and Results

Before performing our main analysis, we ran a preliminary correlation matrix to test the relationship between the raw paper count and our development indicators. This step was critical to determine if we could simply compare total output across nations.

The results revealed an overwhelming bias toward scale. The total number of papers was nearly perfectly correlated with GDP (0.95) and strongly correlated with Population (0.70). In contrast, "quality" metrics like Tertiary Enrollment had much weaker correlations (0.20) with total volume. This confirmed that richer and more populous countries produce more papers simply due to capacity. To identify which countries were actually "performing better" relative to their resources, we needed to normalize the data,so we designed a custom Research Efficiency Metric to penalize "size" and reward "intensity." We defined the final metric as a weighted combination of Z-scores, using the correlations we just found as the weights:

$$\text{Efficiency Metric} = W_{pop} \cdot Z\left(\frac{\text{Papers}}{\text{Population}}\right) + W_{gdp} \cdot Z\left(\frac{\text{Papers}}{\text{GDP}}\right)$$

Where $W_{pop} \approx 0.70$ and $W_{gdp} \approx 0.95$. This formula ensures that a country is ranked based on how much it outperforms the expected output for an economy and population of its size.

To validate the effectiveness of our Efficiency Metric, we compared the global rankings under the traditional "Volume" view against our new "Efficiency" view. The divergence between these two lists confirms that our metric successfully separated scientific success from economic or population mass.

When ranked by raw publication count, the list is predictable. The world's largest economies (China, the United States, and India) dominate the field, producing over 2 million papers combined. However, the color scale (representing our Efficiency Metric) reveals a critical insight: Scale does not equal efficiency. The bars for China and the US are the longest, but their lighter color indicates that their output, while massive, is roughly what we would expect for nations of their colossal size. Their efficiency scores are near zero (average) or even negative.

In contrast, countries like the United Kingdom, Indonesia and Switzerland stand out as they appear in the top tier for volume, yet their darker blue shading shows they also maintain a high efficiency score, outperforming their expected output.

When we rank countries by our Research Efficiency Metric, the list changes completely. The big countries like the US and China disappear, and smaller nations take the top spots. Lebanon is the number one country in this index, followed by Cyprus, Denmark, and Switzerland. The top of the list is very diverse. It includes wealthy European nations like Denmark and Switzerland, but it also includes developing nations like Tajikistan and Burundi. This proves that our formula works as intended. By penalizing GDP and Population, we are not just

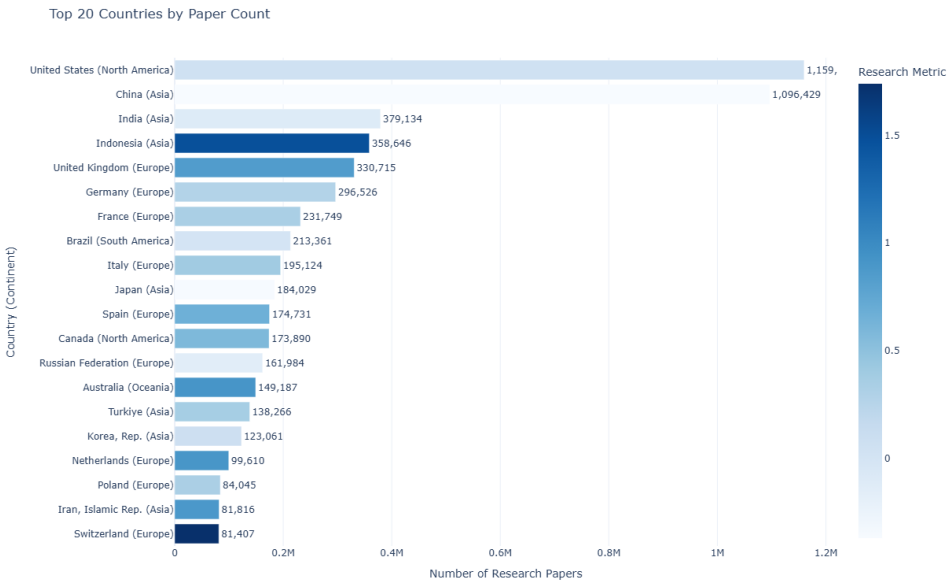

Fig. 2. Top 20 countries ranked by number of papers

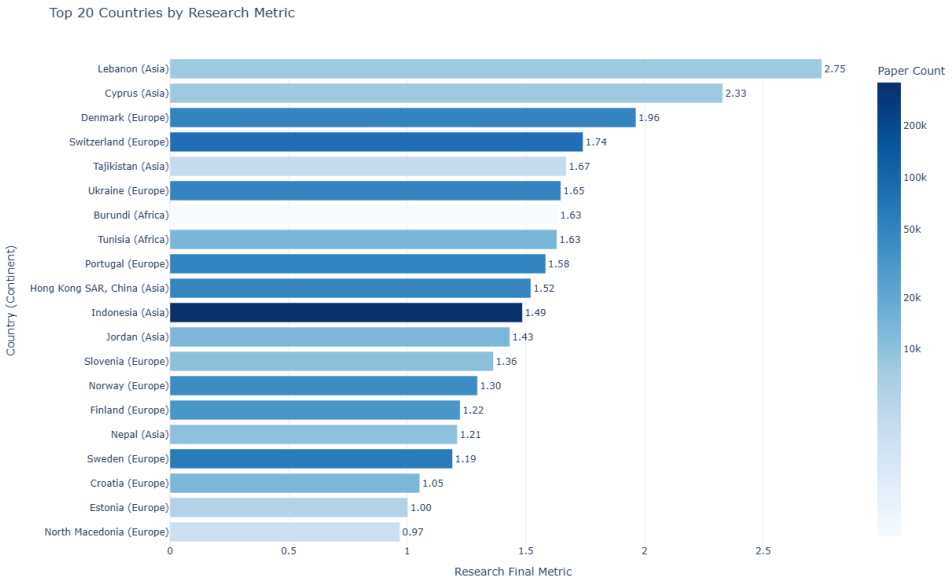

Fig. 3. Top 20 countries ranked by research metric

finding the richest countries. We are finding the countries that produce the most research relative to their specific situation, whether they are wealthy or not.

Topic Specialization: Who Leads in What?

Finally, we looked at each of our three research fields individually to see if the "Efficiency Leaders" change depending on the topic. We compared the Top 20 countries by Volume (Total Papers)(2) against the Top 20 by Efficiency (Our metric)(3) for each field.

The volume list on the left is dominated by the usual countries like China and the United States. However, the labels on these bars reveal their efficiency scores are often very low (e.g. 0.07 for the US), indicating that their dominance is largely a function of their massive size rather than per-capita intensity. Conversely, the efficiency list on the right is populated by "Tech Hubs" like Singapore and Switzerland. While the labels show their total paper counts are a fraction of the giants, their high ranking proves they are achieving outsized results relative to their resources.

The most impressive finding is the group of countries highlighted in orange. These nations:Hong Kong, the United Kingdom, Australia, and the Netherlands, appear in the Top 20 for both volume and efficiency. This proves that it is possible to scale up research production to a global level without sacrificing the high intensity typically associated with smaller specialists.

The analysis of Environmental Science reveals a distinct landscape compared to other fields. When we compare the Top 20 countries by volume against those by efficiency, we see almost no overlap, suggesting a strong divide between the producers of mass research and the most intensive specialists.

The volume list remains consistent with other fields, dominated by massive economies like China and the United States. However, their efficiency scores are relatively low, confirming that their leadership is primarily based on scale rather than per-capita intensity. Conversely, the efficiency ranking reveals a unique mix of leaders. Cyprus takes the top spot, followed by a diverse group that includes Tajikistan, Denmark, Lebanon, Norway, and Nepal. This creates a blend of wealthy Nordic or European nations and developing nations where environmental resource management may be a critical national priority.

Unlike other fields, it is rare for a country to be a leader in both volume and efficiency in Environmental Science. Only two nations appear in both Top 20 lists: Switzerland and Malaysia. These are the only that have managed to achieve massive global scale while maintaining the high per-capita intensity of a specialist.

The results for Medical research are very different from AI or Environment. Here, we see a mix of wealthy European nations and smaller developing countries leading the way in efficiency.

The volume list is standard, with the United States and China producing the most medical research by far. However, their efficiency scores are low, meaning their high output is mostly due to their large size. On the other hand,on the efficiency list Lebanon takes the number one spot, followed by Gambia and Cyprus. This top 20 includes wealthy countries with strong public health systems like Denmark and Switzerland, but also developing nations like Nepal and Tunisia. Five nations: the United Kingdom, Australia, the Netherlands, Switzerland, and Iran appear in both the Top 20 for volume and the Top 20 for efficiency. These nations prove that high volume doesn't have to mean low efficiency.

5 Conclusion

Conclusion

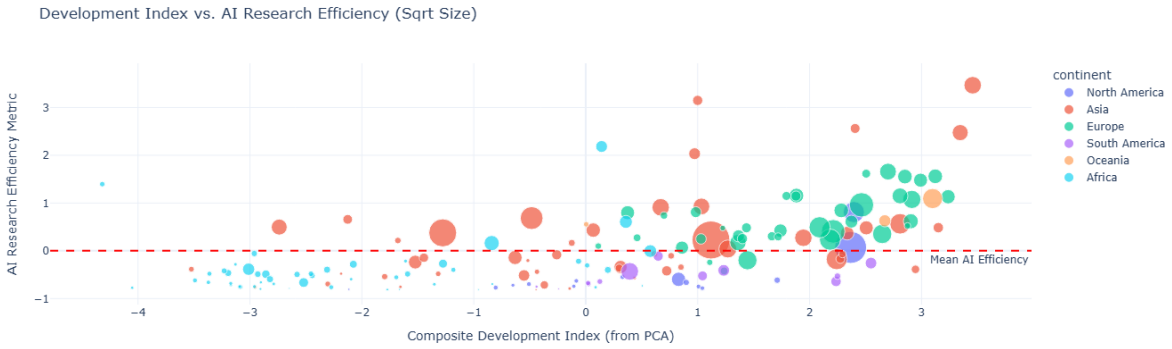

Fig. 4. Development Index vs. AI Research Efficiency

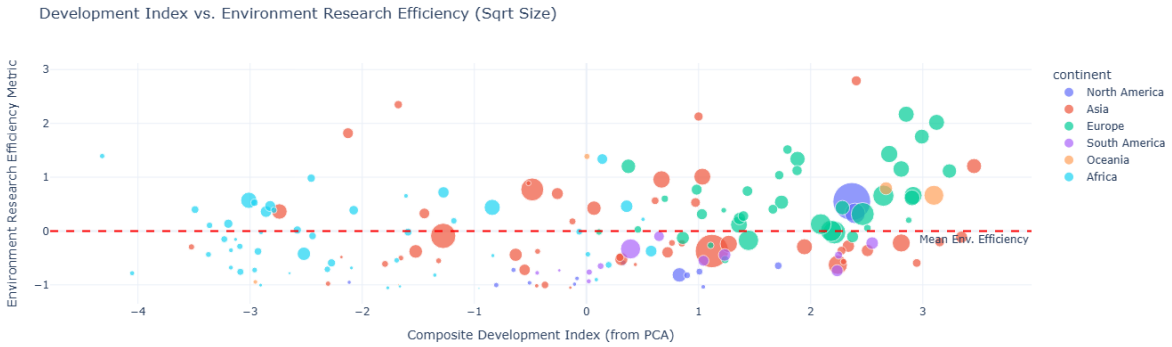

Fig. 5. Development Index vs. Environment Research Efficiency

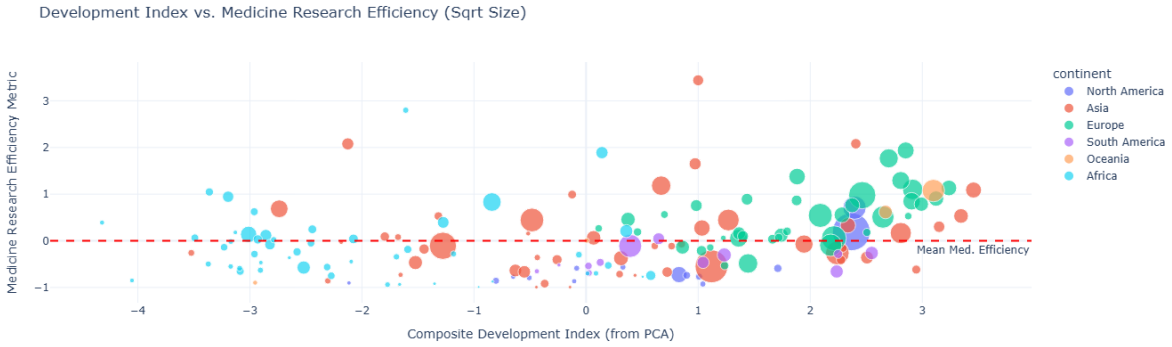

Fig. 6. Development Index vs. Medicine Research Efficiency

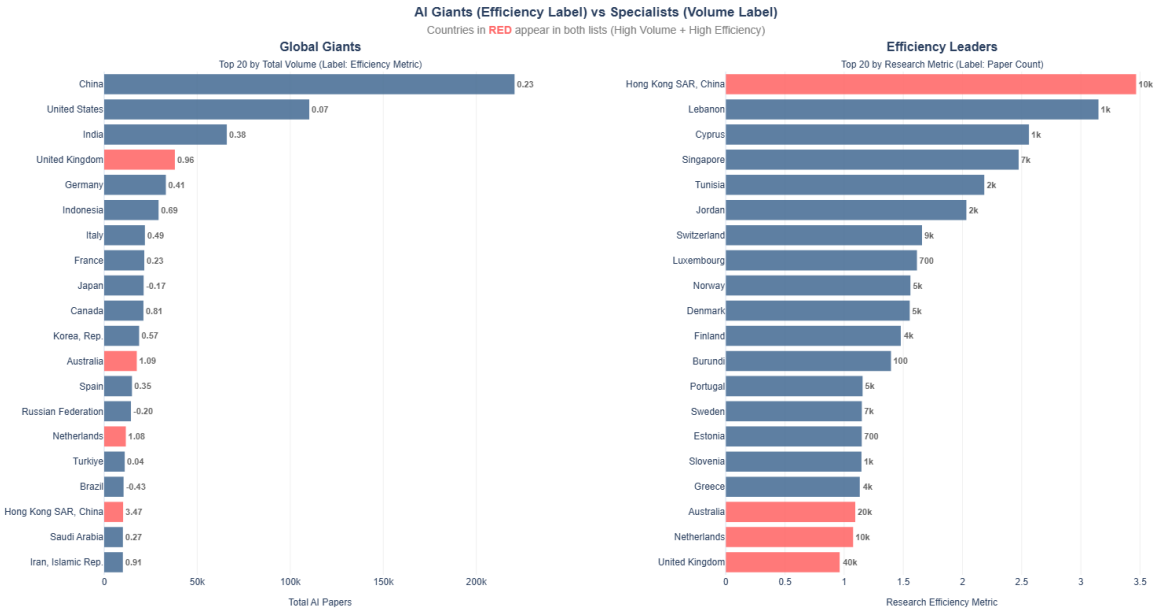

Fig. 7. Artificial Intelligence: Giants vs Specialists

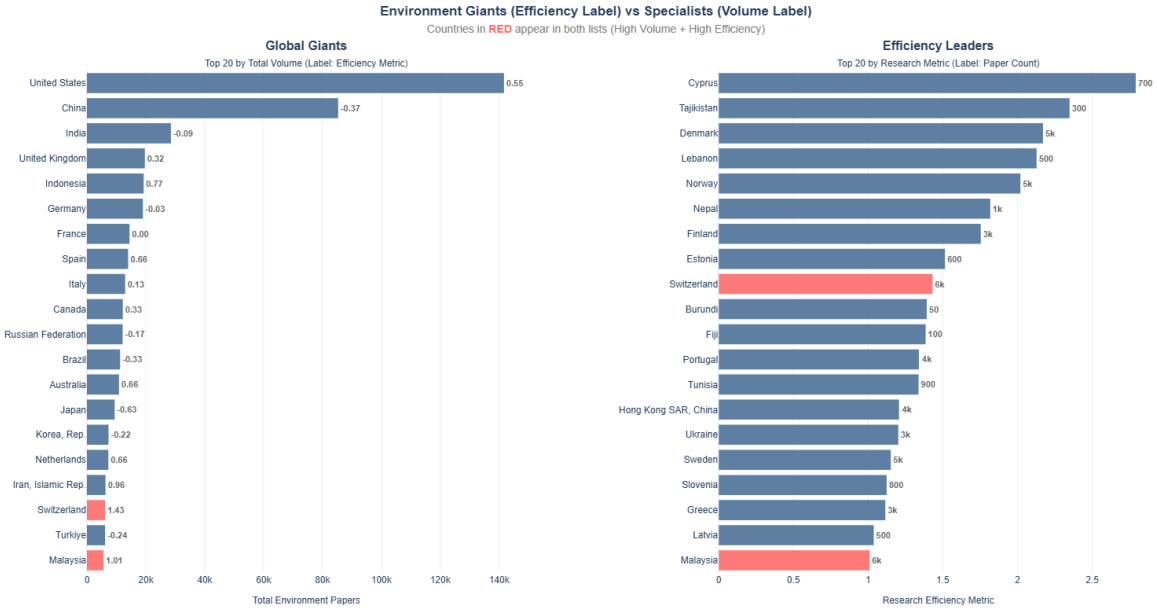

Fig. 8. Environment Research: Giants vs Specialists)

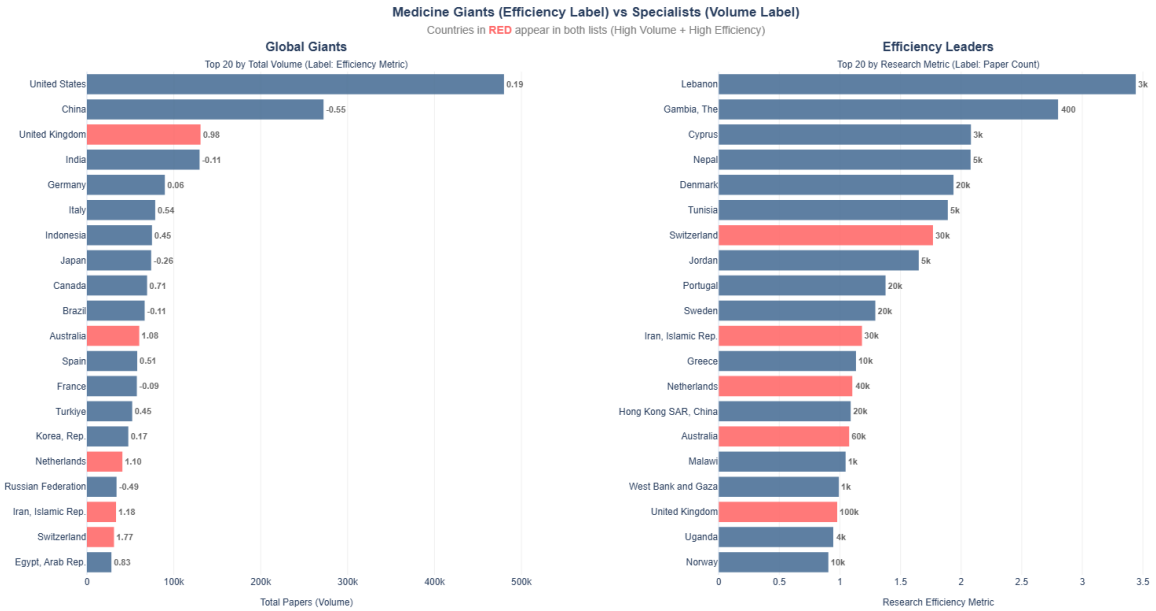

Fig. 9. Medicine: Giants vs Specialists