

Supplementary Materials 1

Data Dictionary and Variable Guide

Available at:

<https://haqueshenas.github.io/lesson-one/scientific-correction-pre-ai/>

OVERVIEW

This document provides a comprehensive description of materials included in *Supplementary Materials 1* and serves as the primary data dictionary for the datasets used in this study.

Supplementary Materials 1 contains three folders:

- **codes**
- **data**
- **univariate_clustering_results**

Codes

The codes folder contains the principal Python programs used to perform the main analyses reported in this study. These scripts represent the core analytical workflows selected from a larger collection of programs that were used during data preparation, country-name harmonization, dataset segmentation, quality control, and production of final tables and figures.

File names beginning with S1, S2, or S3 correspond to the respective sections of the article and indicate the analytical stage for which the code was developed.

Additional methodological information is provided within the script headers and in the associated article. All analytical scripts include a simple graphical user interface that informs users about the required input files. In most cases, the input datasets correspond directly to the CSV files provided in the data folder. Users can therefore execute the scripts directly without modifying the code.

Data

The data folder contains the results and also primary datasets used for the analyses reported in the article.

Three datasets are included:

1. Section_1_Retracton_Rates_By_Country.csv
2. Section_1_All_Retracted_Papers_for_RD_Comparison.csv
3. Sections_2_and_3_dataset.csv

The two Section 1 datasets contain information for 58 countries.

The Sections 2 and 3 dataset contains information for 57 countries because Taiwan was excluded from the economic analyses due to the absence of compatible World Bank indicators required for macro-structural and cross-national economic analyses.

Univariate Clustering Results

The univariate_clustering_results folder contains supplementary clustering outputs generated for individual indicators used throughout the study.

The folder includes:

- Univariate_Clustering_Results.csv
- Publication-quality graphical outputs for each analyzed indicator

For each numerical variable, countries were automatically grouped using Gaussian Mixture Model (GMM) clustering. The optimal number of clusters was determined independently for each indicator using the Bayesian Information Criterion (BIC). Cluster solutions were estimated directly from the original indicator values without prior standardization.

The resulting figures display country rankings, cluster assignments, indicator distributions, and overall means, providing an additional descriptive view of cross-national heterogeneity for individual variables.

These outputs are supplementary and were provided to facilitate transparency, reproducibility, and further exploratory interpretation of the data.

Detailed methodological definitions, indicator construction procedures, normalization methods, weighting schemes, sensitivity analyses, clustering procedures, and statistical analyses are provided in the associated article.

The following sections document the variables contained in the datasets distributed within the data folder.

DATA DICTIONARY STRUCTURE

The remainder of this document provides variable-level documentation for all datasets contained in the *data* folder.

Variable descriptions are organized according to the structure of the original datasets and are presented separately for each CSV file.

The datasets are documented in the following order:

Dataset 1:

Section_1_Retraction_Rates_By_Country.csv

Dataset 2:

Section_1_All_Retracted_Papers_for_RD_Comparison.csv

Dataset 3:

Sections_2_and_3_dataset.csv

For variables appearing in multiple datasets, definitions are provided only once whenever possible in order to avoid redundancy.

Readers interested in methodological details, mathematical formulations, normalization procedures, weighting schemes, clustering algorithms, sensitivity analyses, and statistical modeling strategies can consult the Materials and Methods section of the associated article.

=====

ABBREVIATIONS

- WoS: Web of Science
- RW: Retraction Watch (Retraction Watch Database)
- RR: Retraction Rate
- RD: Retraction Duration

- NRR: Normalized Retraction Rate
- NRD: Normalized Retraction Duration
- NRRDI: Normalized Retraction Rate and Duration Index
- Slope13: Early Correction Concentration Index
- NSlope13: Normalized Slope13
- NRS13I: Normalized Retraction Rate and Slope13 Index
- CI: Confidence Interval
- GDP: Gross Domestic Product
- CPI: Corruption Perceptions Index
- HDI: Human Development Index
- WGI: Worldwide Governance Indicators
- GEE: Government Expenditure on Education
- RDE: Research and Development Expenditure
- AFI: Academic Freedom Index
- GII: Global Innovation Index
- RARE: Retraction-Adjusted Research Efficiency
- ISO3: ISO 3166-1 Alpha-3 country code

DATASET 1

Section_1_Retracton_Rates_By_Country.csv

Description

This dataset provides country-level indicators characterizing national retraction systems using publication volume, retraction frequency, collaboration structure, retraction duration, temporal correction dynamics, and composite normalized indices.

IDENTIFICATION VARIABLES

Country

Country name.

ISO3

ISO 3166-1 Alpha-3 country code.

RETRACTION AND PUBLICATION COUNTS

RW_Single_Count

Number of retracted publications attributed to the country that were classified as non-collaborative (single-country) publications.

RW_Collab_Count

Number of retracted publications attributed to the country that involved international collaboration.

RW_Total_Count

Total number of retracted publications attributed to the country.

WoS_Total_Publications

Total number of publications attributed to the country in the Web of Science dataset used for the analysis.

RETRACTION RATE INDICATORS

RR_Single_%

Retraction rate calculated using only single-country retracted publications.

RR_Collab_%

Retraction rate calculated using internationally collaborative retracted publications.

RR_Total_%

Overall national retraction rate based on all retracted publications.

CONFIDENCE INTERVALS FOR RETRACTION RATE

CI_RR_Total_Lower_%

Lower bound of the confidence interval for the overall national retraction rate.

CI_RR_Total_Upper_%

Upper bound of the confidence interval for the overall national retraction rate.

COLLABORATION STRUCTURE INDICATORS

Collaboration_Share_%

Percentage contribution of internationally collaborative retracted publications to the country's total retraction burden.

CI_Collab_Share_Lower_%

Lower bound of the confidence interval for Collaboration Share.

CI_Collab_Share_Upper_%

Upper bound of the confidence interval for Collaboration Share.

RETRACTION DURATION INDICATORS

The following variables summarize the temporal delay between publication and formal retraction.

Total_Mean_RD_Years

Mean retraction duration (years) calculated using all retracted publications.

Single_Mean_RD_Years

Mean retraction duration (years) calculated using single-country retracted publications.

Collab_Mean_RD_Years

Mean retraction duration (years) calculated using internationally collaborative retracted publications.

RETRACTION DURATION DISTRIBUTION

Total Publications

Total_RD_≤1_%

Percentage of retractions occurring within one year after publication.

Total_RD_1_2_%

Percentage of retractions occurring between one and two years after publication.

Total_RD_2_3_%

Percentage of retractions occurring between two and three years after publication.

Total_RD_3_4_%

Percentage of retractions occurring between three and four years after publication.

Total_RD_4_5_%

Percentage of retractions occurring between four and five years after publication.

Total_RD_>5_%

Percentage of retractions occurring more than five years after publication.

Single-Country Publications

Single_RD_≤1_%

Percentage of retractions occurring within one year after publication.

Single_RD_1_2_%

Percentage of retractions occurring between one and two years after publication.

Single_RD_2_3_%

Percentage of retractions occurring between two and three years after publication.

Single_RD_3_4_%

Percentage of retractions occurring between three and four years after publication.

Single_RD_4_5_%

Percentage of retractions occurring between four and five years after publication.

Single_RD_>5_%

Percentage of retractions occurring more than five years after publication.

Internationally Collaborative Publications

Collab_RD_≤1_%

Percentage of retractions occurring within one year after publication.

Collab_RD_1_2_%

Percentage of retractions occurring between one and two years after publication.

Collab_RD_2_3_%

Percentage of retractions occurring between two and three years after publication.

Collab_RD_3_4_%

Percentage of retractions occurring between three and four years after publication.

Collab_RD_4_5_%

Percentage of retractions occurring between four and five years after publication.

Collab_RD_>5_%

Percentage of retractions occurring more than five years after publication.

TEMPORAL CONCENTRATION INDICATORS

Total_Slope13

Early correction concentration indicator calculated using all retracted publications.

Single_Slope13

Early correction concentration indicator calculated using single-country retracted publications.

Collab_Slope13

Early correction concentration indicator calculated using internationally collaborative retracted publications.

Higher values indicate stronger concentration of retractions during the early post-publication period.

DATA SUFFICIENCY INDICATORS

Total_Valid_RD_Count

Number of valid records used for calculation of Total RD metrics.

Single_Valid_RD_Count

Number of valid records used for calculation of Single RD metrics.

Collab_Valid_RD_Count

Number of valid records used for calculation of Collaborative RD metrics.

Total_RD_Flag_<5

Quality-control indicator showing whether fewer than five valid observations were available for Total RD calculations.

Single_RD_Flag_<5

Quality-control indicator showing whether fewer than five valid observations were available for Single RD calculations.

Collab_RD_Flag_<5

Quality-control indicator showing whether fewer than five valid observations were available for Collaborative RD calculations.

NORMALIZED INDICATORS

All normalized indicators are scaled to the interval [0,1].

Higher values indicate more favorable performance according to the direction specified in the associated article.

NRR_Total_RR

Normalized overall retraction rate.

NRR_Single_RR

Normalized retraction rate for single-country publications.

NRR_Collab_RR

Normalized retraction rate for internationally collaborative publications.

NRD_Total

Normalized mean retraction duration based on all retracted publications.

NRD_Single

Normalized mean retraction duration for single-country publications.

NRD_Collab

Normalized mean retraction duration for internationally collaborative publications.

NRRDI_Total_RR

Composite indicator integrating normalized retraction rate and normalized retraction duration for all retracted publications.

NRRDI_Single_RR

Composite indicator integrating normalized retraction rate and normalized retraction duration for single-country publications.

NRRDI_Collab_RR

Composite indicator integrating normalized retraction rate and normalized retraction duration for internationally collaborative publications.

NSlope13_Total

Normalized early correction concentration indicator based on all retracted publications.

NSlope13_Single

Normalized early correction concentration indicator based on single-country publications.

NSlope13_Collab

Normalized early correction concentration indicator based on internationally collaborative publications.

NRS13I_Total_RR

Composite indicator integrating normalized retraction rate and normalized early correction concentration for all retracted publications.

NRS13I_Single_RR

Composite indicator integrating normalized retraction rate and normalized early correction concentration for single-country publications.

NRS13I_Collab_RR

Composite indicator integrating normalized retraction rate and normalized early correction concentration for internationally collaborative publications.

DATASET 2

Section_1_All_Retracted_Papers_for_RD_Comparison.csv

Description

This dataset contains article-level records used for retraction-duration analyses and cross-country comparisons presented in Section 1.

Record ID

Unique article identifier assigned in the Retraction Watch database.

Country

Country attributed to the retracted publication.

ISO3

ISO 3166-1 Alpha-3 country code.

Type

Collaboration category of the retracted publication.

Possible values:

Single = single-country publication

Collab. (Collaboration) = internationally collaborative publication

RD_Years

Retraction duration expressed in years.

Represents the elapsed time between publication and formal retraction.

DATASET 3

Sections_2_and_3_dataset.csv

Description

This dataset combines the scientific-correction indicators generated in Dataset 1 with the socioeconomic, governance, innovation, academic freedom, and economic-accounting variables used in Sections 2 and 3.

Several variables appearing in this dataset are identical to those previously defined in Dataset 1. Readers should refer to the corresponding definitions above for those variables.

SOCIOECONOMIC AND INSTITUTIONAL VARIABLES

GDP (log_2000_2024)

Natural logarithm of the average national Gross Domestic Product during 2000–2024.

CPI (2000_2024)

Average Corruption Perceptions Index during 2000–2024.

Higher values indicate lower perceived corruption.

RDE (2000_2023)

Average Research and Development Expenditure during 2000–2023.

Expressed as percentage of GDP.

GEE (2000_2024)

Average Government Expenditure on Education during 2000–2024.

HDI (2000_2023)

Average Human Development Index during 2000–2023.

WGI (2000_2024)

Average Worldwide Governance Indicators score during 2000–2024.

AFI (2000_2024)

Average Academic Freedom Index during 2000–2024.

GII (2000_2025)

Average Global Innovation Index during 2000–2025.

ECONOMIC ACCOUNTING INDICATORS**Gross_R&D_\$**

Estimated gross national expenditure on research and development before adjustment for retraction exposure.

Net_R&D_\$

Estimated R&D investment remaining after adjustment for observed retraction exposure.

Net_R&D_%

Retained research-investment intensity after accounting for retraction exposure.

Retraction_Cost_\$

Estimated economic cost associated with retracted research output.

Retraction_Cost_%

Percentage of national R&D investment associated with retracted publications.

Paper_Cost_\$

Estimated R&D investment per published paper.

Cost_per_Retracted_Paper_\$

Estimated economic cost associated with each retracted publication.

RARE

Retraction-Adjusted Research Efficiency.

Represents valid scientific output generated per unit of estimated national R&D expenditure after accounting for retracted publications.

Higher values indicate greater research efficiency.

LOG-TRANSFORMED VARIABLES

Gross_R&D_\$_log

Log-transformed Gross R&D Expenditure.

Retraction_Cost_\$_log

Log-transformed Retraction Cost.

Paper_Cost_\$_log

Log-transformed Paper Cost.

Cost_per_Retracted_Paper_\$_log

Log-transformed Cost per Retracted Paper.

RARE_log

Log-transformed Retraction-Adjusted Research Efficiency.

Net_R&D_%_log

Log-transformed Net R&D Expenditure (%).

Log transformation was applied to improve cross-country comparability, reduce skewness, and enhance statistical stability.

DISPLAY AND VISUALIZATION VARIABLES

GDP_BillionUSD

Gross Domestic Product expressed in billion USD.

Used primarily for visualization and interpretation.

Retraction_Cost_BillionUSD

Retraction Cost expressed in billion USD.

Used primarily for visualization and interpretation.

MULTIVARIATE STRUCTURAL CLASSIFICATION VARIABLES

Economic_Cluster

Economic-retraction regime membership identified using hierarchical agglomerative clustering.

Clusters were constructed using four standardized dimensions:

- Retraction Rate (RR)
- Research and Development Expenditure (RDE)
- Cost per Retracted Publication
- Retraction-Adjusted Research Efficiency (RARE)

Countries assigned to the same cluster exhibit similar economic and scientific-correction structures.

Cluster numbers are categorical identifiers and do not imply ranking.

Hierarchical_Clusters

Multidimensional structural regime classification obtained from hierarchical clustering performed in the integrated economic-retraction feature space.

The clustering solution was derived from variables representing:

- Retraction Rate (RR)
- Net R&D Expenditure
- Retraction Cost
- Retraction-Adjusted Research Efficiency (RARE)

Cluster assignments identify countries sharing similar structural configurations in the political economy of scientific correction systems.

Cluster labels are nominal categories and should not be interpreted as ordinal rankings.

NOTES

1. All percentages are expressed on a 0–100 scale unless otherwise noted.
2. Retraction duration variables are expressed in years.
3. Normalized variables are expressed on a 0–1 scale.
4. Confidence intervals are reported on the same scale as the corresponding indicator.
5. Countries with very small numbers of valid retraction-duration observations should be interpreted cautiously; the corresponding RD_Flag_<5 variables identify such cases.
6. Cluster identifiers are nominal labels and do not represent rankings or performance scores.
7. This data dictionary provides concise operational descriptions of variables. Detailed mathematical formulations, normalization procedures, weighting schemes, sensitivity analyses, cluster-construction methods, and theoretical justification are reported in the Methods section of the associated article.