Data Science and Global Health: investigating the relationship between disease with high burdens and data science methodologies

SUDS Scholar: Bilin Nong, University of Toronto

Supervisor: Prof. Aya Mitani, Dalla Lana School of Public Health, University of Toronto

Introduction

- The landscape of global healthcare is rapidly transforming, driven by increasing disease burdens and novel data science methodologies.
- Disease burden, measured by metrics such as Disability-Adjusted Life Years (DALYs), is an indicator of population health[4].
- Understanding disease burdens is crucial for effective public health planning and resource allocation.
- The rise of Big Data Analytics[2, 3] offers opportunities to enhance both patient care and health management systems.
- While novel data science methodologies are often motivated by research studies related to diseases, less is known about the relationship between common disease burdens and methodology research motivated by them.

Objectives

- Assess the influence of trending global diseases on data science methodologies through a systematic literature review.
- Identify diseases that may be overlooked by methodological research.

Methods

Data Preparation

- We compiled a dataset for top 25 diseases that are leading causes of global DALYs in 3 age groups (all ages, age0-9, age75+)[5], encompassed disease names, DALY rankings, and disease types.
- Conducted literature search on Web of Science (WoS) Database, limiting the scope to publications between 2010 and 2024 in WoS data science category. The selected articles were further filtered by the data science journal list from NYU[6].
- Linked each disease with corresponding methodological research articles information, and ranked all diseases based on total number of relevant publications.

Statistical Analysis

- We calculated the **Spearman's rank-order correlation coefficient** between disease DALY rankings and rankings of associated publication[1]:

$$\varrho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where x_i is the DALY ranking of i^{th} ranked disease; y_i is the publication ranking of i^{th} ranked disease; $\bar{x}$ and $\bar{y}$ is the mean of all DALY rankings and publication rankings.

- We also utilized **Mann-Kendall Trend Test** to test for monotonic trend in total numbers of relevant publications over time for each disease.
- Created a **heatmap** to visualize the relationship between diseases and journals with articles motivated by them.

Results

	Articles (N=41831)
Publication Year	
2010-2014	9748 (23.3%)
2015-2019	12885 (30.8%)
2020-2024	19198 (45.9%)
Times Cited	
Mean (SD)	15.2 (128)
Median [Min, Max]	4.00 [0, 17700]
Journal	
Statistics in Medicine	2067 (4.9%)
Biometrics	848 (2.0%)
Statistical Methods in Medical Research	746 (1.8%)
Applied Mathematics and Computation	575 (1.4%)
Risk Analysis	555 (1.3%)
Others	37040 (88.5%)
Country of Affiliation	
USA	10487 (25.1%)
Peoples R China	6668 (15.9%)
India	2024 (4.8%)
England	1661 (4.0%)
Canada	1342 (3.2%)
Other Countries	19649 (47.0%)
Disease Type	
Non-communicable Disease	22669 (54.2%)
Communicable Disease	18569 (44.4%)
Injury	593 (1.4%)
Computer Science Category	
Yes	5148 (12.3%)
No	35507 (84.9%)
Missing	1176 (2.8%)
Statistics Category	
Yes	14679 (35.1%)
No	25976 (62.1%)
Missing	1176 (2.8%)
Mathematics Category	
Yes	34791 (83.2%)
No	5864 (14.0%)
Missing	1176 (2.8%)

Table 1. Summary of characteristics of collected articles.

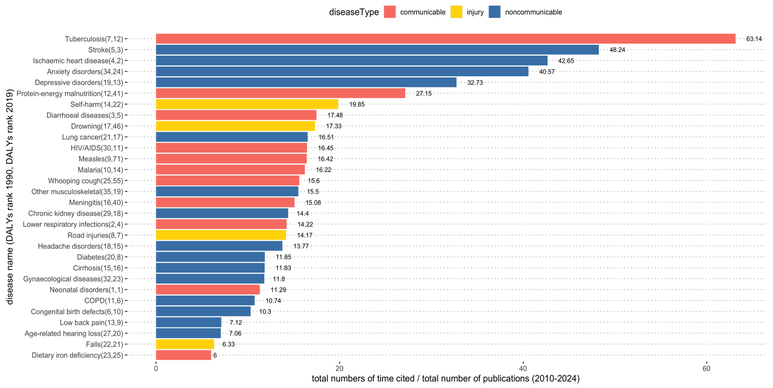

Figure 1. Bar plot showing number of citations versus publications ratio for top diseases in all ages.

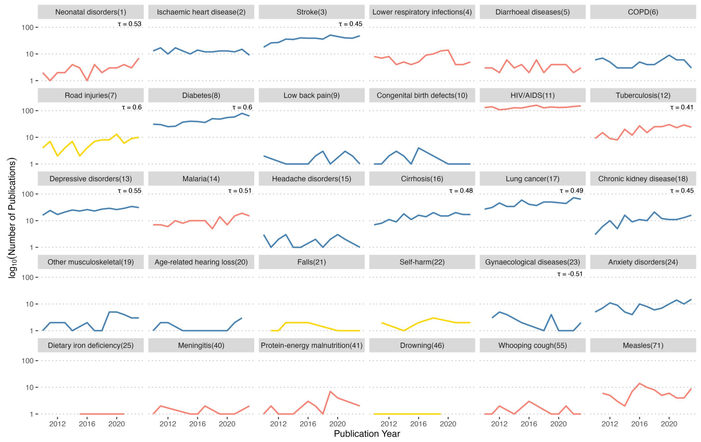

Figure 2. Line chart of $\log_{10}$(total publications) between 2010 and 2023 for top diseases in all ages. The Kendall's τ with p-value < 0.05 in trend test are reported.

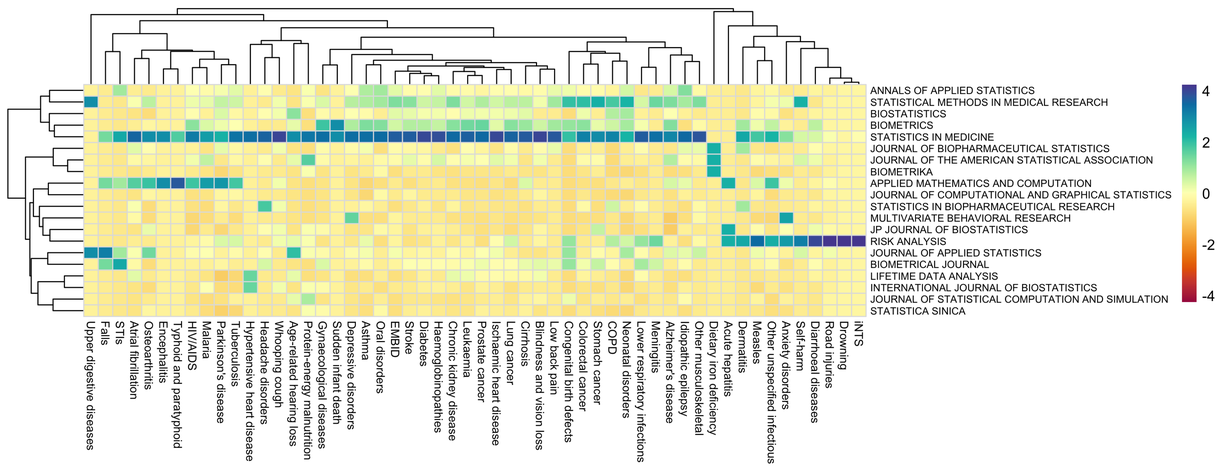

Figure 3. Heatmap of total number of publications aggregated by each disease and journal. Cell values were scaled and centered in the column direction. Rows and columns are clustered using hierarchical clustering method.

Summary of Results

- No. of data science publications related to top diseases are increasing, WoS categorized **83%** of articles into Math (Table 1).
- Tuberculosis has the highest citations vs. publications ratio (**63.14**), the DALY ranking went down 5 places from 1990 to 2019. However, although neonatal disorder is the top leading cause of global DALY in 1990 and 2019, the ratio is only **11.29** (Figure 1).
- Statistical analysis demonstrated a weak correlation between DALY rankings and publication quantities across all age groups (Spearman's correlation = **0.34**).
- No. of data science publications motivated by HIV/AIDS is consistently high. Publications on depressive disorder, diabetes, malaria, neonatal disorder, and road injuries are increasing. Publications on gynecological diseases are decreasing (Figure 2) .
- Annals of Applied Statistics, Statistical Methods in Medical Research, Biostatistics, Biometrics, and Statistics in Medicine tend to publish data science methodology articles motivated by similar diseases. Risk Analysis tends to publish articles on injuries and communicable diseases (Figure 3).

Conclusion

- Significant disparities exist between disease DALY rankings and research focus, with some high-burden diseases receiving disproportionately less attention in methodological research.
- Highlights a potential misalignment between global health priorities and current research focus in data science.
- A more balanced research focus may help researchers contribute effectively to improving the public health planning process.

Bibliography

1. Akoglu, Haldun. 2018. "User's Guide to Correlation Coefficients." Turkish Journal of Emergency Medicine 18 (3): 91–93.
2. Batko, Kornelia, and Andrzej Ślęzak. 2022. "The Use of Big Data Analytics in Healthcare." Journal of Big Data 9 (1): 3.
3. Corsi, Alana, Fabiane Florencio De Souza, Regina Negri Pagani, and João Luiz Kovaleski. 2021. "Big Data Analytics as a Tool for Fighting Pandemics: A Systematic Review of Literature." Journal of Ambient Intelligence and Humanized Computing 12 (10): 9163–80.
4. Udompap, Prowpanga, Donghee Kim, and W. Ray Kim. 2015. "Current and Future Burden of Chronic Nonmalignant Liver Disease." Clinical Gastroenterology and Hepatology 13 (12): 2031–41.
5. Vos, Theo, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, et al. 2020. "Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019." The Lancet 396 (10258): 1204–22.
6. Statistics and Probability Journals (NYU) . https://pages.stern.nyu.edu/~jss2/statjournals.html。

Acknowledgement

UNIVERSITY OF TORONTO | DATA SCIENCES INSTITUTE