MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230973

Analyzing the Spread of Informatics with PubMed

Brian E. CHAPMAN[a,1], Wendy W. CHAPMAN[b] and Jeremiah CHAPMAN[c]
[a]*Melbourne Medical School, University of Melbourne, VIC, Australia*
[b]*Centre for Digital Transformation of Health, University of Melbourne, VIC, Australia*
[c]*Australian National University, ACT, Australia*
ORCiD ID: Brian E. Chapman https://orcid.org/0000-0003-0815-5448, Wendy Chapman https://orcid.org/0000-0001-8702-4483, Jeremiah B. Chapman https://orcid.org/0000-0003-3023-6287

Abstract. We analyzed PubMed citations since 1988 to explore the dissemination of medical/health informatics concepts between countries and across medical domains. We extracted countries from the PubMed author affiliation field to identify and analyze the top 10 informatics publishing countries. We found that the informatics publications are becoming more similar over time and that the rate of exchange across countries has increased with the introduction of e-publishing. Nonetheless, with the exception of machine learning, the impact of core informatics concepts on mainstream medicine and radiology publications remains small.

Keywords. MeSH, bibliometrics, medical informatics

1. Introduction

Bibliometric analysis can provide insight into a domain and facilitate comparisons across fields and countries. Fan et al. [1] used Web of Science to characterize world productivity in the field of minimally invasive spine surgery, highlighting temporal and geopolitical dynamics. Similarly, Farhat et al. [2] compared congenital heart disease publication rates in Arab versus developed countries. We use bibliometrics to create a portrait of health informatics to elucidate its evolving concept of self and relationship to the broader medical community. Health Informatics is an intellectual discipline that arose in the 1960s as researchers began using analog and digital computers to study the nature and use of medical information while simultaneously using computers to shape the delivery of healthcare. Informatics researchers respond to the challenges in the specific locales where they work, thus the health informatics literature is likely culturally influenced, depending on how countries organize their healthcare. To understand how medical informatics has evolved over time and space, we used MeSH tags in the literature to characterize its changes over time and across domains and countries.

[1] Corresponding Author: Brian E. Chapman, email: brian.chapman@unimelb.edu.

2. Methods

The 2021version PubMed database was downloaded from the National Library of Medicine (USA) (ftp.ncbi.nlm.nih.gov). The XML files were parsed and entries for a subset of Journals from the domains of 1) health informatics, 2) general medicine, and 3) radiology were extracted. For health informatics journals we used the 93 journals listed in the NLM Broad Subject Terms as Medical Informatics. For general medicine, we used BMJ, the Journal of the American Medical Association, The Lancet, the New England Journal of Medicine, and Annals of Internal Medicine. For radiology we used Radiology, European Radiology, AJR American Journal of Roentgenology, European Journal of Radiology, and Investigative Radiology. The 2021AB version of MeSH was downloaded from Bioportal (https://bioportal.bioontology.org/) and used for all our analyses.

For this preliminary analysis we used cosine similarity based on MeSH tags to measure similarity between journals, disciplines, and countries. One-hot encodings (vocabularies comprising a flat list of all terms in the sub-tree) were defined for all of MeSH and for the Information Science subset of MeSH. This simplification discards the hierarchical knowledge contained in MeSH, which will result in an underestimation of actual similarity. We counted the number of times a MeSH term was used to tag articles. MeSH annotations include major and non-major tags; a major tag represents a central concept of the article, as identified by the NLM annotators. We previously found that limiting our analysis to major tags better elucidated research trends within a discipline, so we focus on major terms only. We aggregated these counts using the MeSH graph structure, so that for each MeSH term we had both the number of citations for that term and the number of citations for all its descendants, which we typically lumped together.

To compare MeSH usage across countries, we first needed to identify the countries where the authors resided at the time of publication using the affiliation field which was added to Medline in 1988. The data, for the most part, are provided directly from the publisher, with no quality control by the NLM. Consequently, the data are extremely variable and noisy. We developed a regular expression-based multi-step filter for affiliation detection. Data for the filter were drawn primarily from two sources. First, a list of universities for each country was obtained from https://github.com/endSly/world-universities-csv and modified and simplified to match this task (updated list is available in our GitHub repository). Country names as well as 2- and 3-character ISO country abbreviations, and US state names and abbreviations were drawn from the Python package geonamescache. We deleted unpopulated regions and added common but non-ISO abbreviations (e.g. “UK”, “U.S.A.”) and states and provinces for Canada and Australia. Finally, we created a list of health systems, hospitals, government agencies, and research institutions by viewing samples of PubMed citations.

Our filter has six consecutive steps. It proceeds to the next step only if it fails to find a country in the current step. The six steps are as follows: 1) university names, 2) e-mail addresses, 3) country names, 4) ISO-3 country codes, 5) state/province names, and 6) ISO-2 country codes. To evaluate the performance of the Detector, a co-author manually analyzed a random sample of 200 articles. The manually assigned countries were then compared to the automated classifications. We found 218 true positives, one false positive, and 22 false negatives (some articles had authors from multiple countries) for accuracy of 92.6%, recall of 90.8%, and precision of 99.5%.

Time lags exist between a MeSH term's first occurrence in different countries. We hypothesized that this lag time has decreased with the advent of electronic publication. We used the Medline Article *PubModel* field to count the number of electronic publications per year and then computed the ratio of e-publications to all publications (Fig. 1). Based on the graph, we defined a pre-e-publication era as before 2000 and the post-e-publication era as starting in 2006, leaving 2000-2006 as a transition phase. To examine the impact of e-publication on the lag time between countries, we selected Information Science terms that were introduced between 1990 and 2000 for which there were at least a cumulative of 25 US publications by 2021. Terms for this set were dropped if there were no other countries with at least 15 publications with that term. A similar process was followed for identifying post-e-publication terms introduced between 2006 and 2016. This process resulted in 38 pre-terms and 20 post-terms. For each country and term we identified the lag between when the first country and subsequent countries reached a cumulative publication of 1 paper per ten million residents.

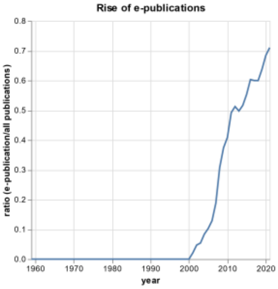

Figure 1. The ratio of e-publications to all publications by year of publication.

3. Results

We identified the top 10 countries publishing informatics articles as US, GB, CA, DE, CN, AU, FR, NL, IT, and JP. Figure 2 shows the cosine similarity for these countries (relative to the US). Both graphs show medical informatics publications are becoming more similar over time.

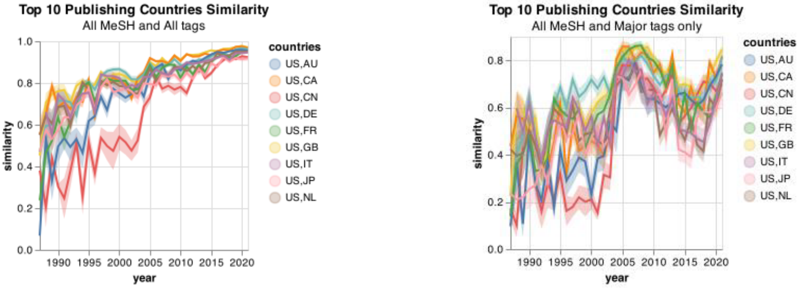

Figure 2. Medical informatics cosine similarity for the top 10 informatics publishing countries. Looking at all MeSH tags (left), we see an increasing overall similarity between countries. Looking at major tags only we see a much more complex dynamic, with similarity decreasing in our initial post-e-publication era and then becoming more similar after 2015, which we might label as the "machine learning era."

Figure 3 illustrates the concept of lag time. The term "Vocabulary, Controlled" was introduced in 1996 and (as of 2021) has been assigned to 1,713 total articles from US authors (maximum) and 36 total articles from JP authors (minimum). "Machine Learning", introduced in 2016, had a maximum number of 1,016 (US) and minimum of 54 (NL) tag assignments. The left and center graphs in Figure 3 show the normalized cumulative sums for publications in medical informatics reference. The 38 pre- MeSH terms and 22 post- MeSH terms that we analyzed for between-country informatics lag times showed a decrease from 6.6 years to 3.7 years lag time. A Mann-Whitney U test showed this to be statistically significant (U=688.5, two-sided p-value: <0.001).

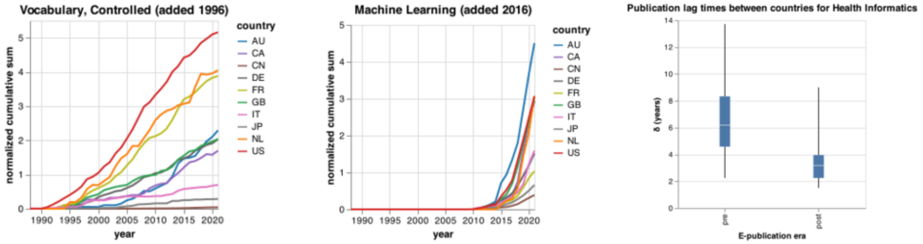

Figure 3. Left and middle show example normalized cumulative publication rate trends by country. "Vocabulary, Controlled" was introduced in the pre-e-publication era, whereas "Machine Learning" was a post-e-publication term. Right shows a box and whisker plot for between-country lag times (δ in years) for all pre -and post-e-publication MeSH terms.

Figure 4 shows normalized yearly aggregate tag counts for Information Science terms in medical informatics, medicine, and radiology journals, illustrating the dominance of Computing Methodologies and Informatics in medical informatics and radiology journals and Communication and Information Management in medicine, with less diversity overall in radiology.

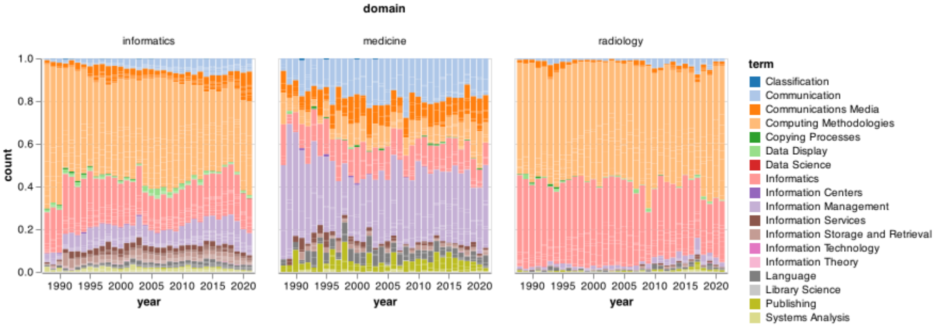

Figure 4. Normalized yearly aggregate tag counts for level 1 Information Science terms.

We examined the influence of the subset of terms studied in the lag analysis, which includes children of the terms in Fig. 4, by plotting the most frequently tagged MeSH terms from this subset in medicine and radiology (Fig. 5). It is notable in the left graph (medicine) how infrequently the core concepts of our field occur, such as "Medical Records Systems, Computerized" (n=179) and "Decision Support Techniques" (n=363). In radiology, "Machine Learning" and "Neural Networks, Computer" dominate.

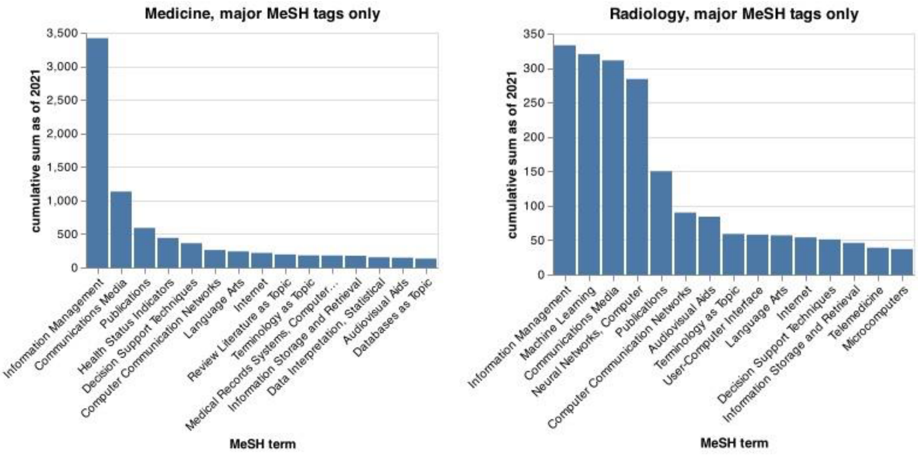

Figure 5. Cumulative publication numbers to 2021 for the top 15 MeSH terms analyzed in the e-publication lag analysis (both pre- and post-epublication) for medicine (left) and radiology (right).

4. Discussion

Analysis of MeSH tags provided insight into the impact of the internet on the spread of informatics concepts. Limitations include an imperfect affiliation analysis, the fact that the analysis of concept spread across countries may be biased by US-centric aspects of MeSH, and our relatively simplistic treatment of the MeSH hierarchy, for example by using cosine similarity to compare country similarity in Fig. 2.

5. Conclusions

This analysis was a hermeneutical portrait of medical informatics. Through MeSH analysis, we have seen informatics concepts spread widely around the world with e-publications increasing the speed at which these concepts diffuse across countries. Yet, with few exceptions, such as "Machine Learning" in radiology, medical informatics impact on mainstream medical publications remains limited.

References

[1] Fan G, Han R, Zhang H, He S, Chen Z. Worldwide research productivity in the field of minimally invasive spine surgery: a 20-year survey of publication activities. Spine (Phila Pa 1976). 2017 Nov;42(22):1717-22, doi: 10.1097/BRS.0000000000001393.

[2] Farhat T, Abdul-Sater Z, Obeid M, Arabi M, Diab K, Masri S, Al Haless Z, Nemer G, Bitar F. Research in congenital heart disease: a comparative bibliometric analysis between developing and developed countries. Pediatr Cardiol. 2013 Feb;34(2):375-82, doi: 10.1007/s00246-012-0466-6.