BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open

Improving the evaluation of worldwide biomedical research output: classification method and standardized bibliometric indicators by disease

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-020818
Article Type:	Research
Date Submitted by the Author:	24-Nov-2017
Complete List of Authors:	van de Laar, Lissy; Gupta Strategists, de Kruif, Thijs; Gupta Strategists Waltman, Ludo; Leiden University, Centre for Science and Technology Studies Meijer, Ingeborg; Universiteit Leiden, CWTS Gupta, Anshu; Gupta Strategists Hagenaars, Niels; Gupta Strategists
Keywords:	HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS, HEALTH ECONOMICS

SCHOLARONE™
Manuscripts

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

IMPROVING THE EVALUATION OF WORLDWIDE BIOMEDICAL RESEARCH OUTPUT: CLASSIFICATION METHOD AND STANDARDIZED BIBLIOMETRIC INDICATORS BY DISEASE

Corresponding author
Lissy van de Laar, MSc
Gupta Strategists, PO Box 16, 4060 GA Ophemert
lissy.vandelaar@gupta.nl
0031 6 34 59 35 07

Co-authors
Ir. Thijs de Kruif, Gupta Strategists, The Netherlands
Dr. Ludo Waltman, Centre for Science and Technology Studies, Leiden University, The Netherlands
Dr. Ingeborg Meijer, Centre for Science and Technology Studies, Leiden University, The Netherlands
Dr. Anshu Gupta, Gupta Strategists, The Netherlands
Dr. Niels Hagenaars, Gupta Strategists, The Netherlands

Key words
Bibliometrics [MeSH], Data mining [MeSH], Classification [MeSH]

Word count excluding title page, abstract, references, figures and tables
3273 words

For peer review only
CONFIDENTIAL

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

ABSTRACT

Objective: Since most biomedical research focuses on a specific disease, evaluation of research output requires disease-specific bibliometric indicators. Currently used methods are insufficient. The aim of this study is to develop a method that enables detailed analysis of worldwide biomedical research output by disease.

Design: We applied text mining techniques and analysis of author keywords to link publications to disease groups. Fractional counting was used to quantify disease-specific biomedical research output of an institution or country. We calculated global market shares of research output as a relative measure of publication volume. We defined `top publications` as the top 10% most cited publications per disease group worldwide. We used the percentage of publications from an institution or country that were top publications as an indicator of research quality.

Results: We were able to classify 54% of all 6.5 million biomedical publications in our database (based on Web of Science) to a disease group. We could classify 78% of these publications to a specific institution. We show that between 2000 and 2012 'Other infectious diseases' was the largest disease group with 337,485 publications. Lifestyle diseases, cancers, and mental disorders have grown most in research output. The USA was responsible for the largest number of top 10% most cited publications per disease group, with a global share of 45%. Iran (+3,500%) and China (+700%) have grown most in research volume.

Conclusions: The proposed method provides a tool to assess biomedical research output in new ways. It can be used for evaluation of historic research performance, to support decision making in management of research portfolios, and to allocate research funding. Furthermore, using this method to link disease-specific research output to burden of disease can contribute to a better understanding of the societal impact of biomedical research.

STRENGTHS AND LIMITATIONS OF THIS STUDY

Strengths

- The proposed method offers quantitative insight in research quantity and quality for 269 disease groups.
- The proposed method can be used for evaluation of historic research performance at disease level. It can support decision making in management of research portfolios, showing relative strengths and weaknesses of institutions and countries, as well as identifying research gaps at national and global level. It can also be valuable in allocation of research funding.

Limitations

- Author keywords were used instead of the standardised MeSH descriptors, which are not available in the Web of Science database.
- Research about for instance molecular mechanisms, medical techniques, and health sciences could often not be classified to a specific disease group and was thus not included in our results.

CONFIDENTIAL For peer review only

INTRODUCTION

One of the goals of biomedical research is to eradicate burden of disease. The grand societal challenges in European funding also build on the premise that (biomedical) research should contribute to prevention and treatment of diseases [1].

Yet surprisingly, biomedical research output has not been systematically catalogued by diseases so far [2]. Most publicly available metrics for analysing biomedical research by topic have severe limitations. Research fields in the Web of Science database produced by Clarivate Analytics are defined at a too high level, since they cover a complete medical specialism [3]. The Scopus database produced by Elsevier has the same problem. Medical Subject Headings (MeSH) terms [4] are more specific, but are available only for a selection of journals.

Several authors have made efforts to analyse research output and funding at disease level, but only for a selection of diseases. Evans et al compared research output between countries for 19 disease groups, based on the International Classification of Disease (ICD)-9 chapters [5]. Gillum et al [6] and Gross et al [7] analysed burden of disease and research funding for a selection of 29 conditions, derived from the ICD. In various other studies funding, research output and burden of disease were described for specific diseases in a case by case approach. This was done for example for yellow fever [8] and neglected tropical diseases [9]. In other studies, total biomedical research output was analysed for specific countries [10, 11] or compared between countries [12, 13].

Text mining techniques are increasingly applied to biomedical text to uncover unseen relationships [14]. In this study we use these techniques to create a reference structure of disease groups and to catalogue publications accordingly. This opens a bridge between biomedical research output and other information available at disease level, which can contribute to a better understanding of the societal impact of biomedical science.

METHODS

Selection of biomedical publications

The analysis was based on the Clarivate Analytics WoS database available at the Centre for Science and Technology Studies (CWTS) of Leiden University. Since the goal of this study is to quantify research output by disease, we included biomedical research fields only. Of the 250 WoS research fields, we selected the 84 fields that are most medically oriented. We validated the selection by looking at the research output of the eight Dutch university medical centres: over 98% of their publications were in one of these fields. Appendix 1 provides a full list of research fields included in this study. The dataset was compiled in June 2014. It includes all publications in the 84 selected research fields, published between 2000 and early 2014, with WoS document type ‘article’ or ‘review’. Not all publications from the first six months of 2014 were available, due to periodical updating of the CWTS in-house version of the WoS database. The dataset contained 6.5 million publications in total.

Classification of publications by disease group

We defined 269 disease groups, based on the ICD-10 classification and covering the full spectrum of this classification. We used a two-step approach to categorise publications to disease groups.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

First, we categorised the author keywords listed by authors in their publications. In total, 158,700 unique author keywords were used in at least ten publications in our dataset. Of these keywords, the 32,400 most frequently used keywords (used in more than 70 publications each) were short listed and further evaluated. 21% of these keywords were specific for a single disease group. For example, the keyword 'Alzheimer's disease' was linked to 'dementia'. Many keywords were not suitable to use for categorisation to disease groups because they were either too general or not disease-specific. Examples of keywords not linked to a disease group are 'inflammation' and 'keyhole surgery'. We note that not all publications include author keywords.

In the second step, a text mining algorithm was used to search for disease-specific terms in titles and abstracts of publications. In this step, first a list of 10,983 unambiguous, disease-specific terms was generated by hand by medical professionals to characterize specific disease groups. Examples of terms for the disease group 'malignant neoplasm prostate' include 'prostate cancer', 'prostate carcinoma', 'malignant tumor prostate', and 'sarcoma prostate'. The generated disease-specific terms were then reviewed by another medical professional for ambiguity. Subsequently publications with one of these 10,983 terms in either title or abstract were assigned to the corresponding disease group. If the same publication was assigned to multiple disease groups, it was fully counted for all of them.

The method was validated in several ways. The first step was a manual examination of a random sample of 680 publications assigned to a disease group. Subsequently, a random sample of 315 publications not assigned to a disease group was manually examined. The examination was executed by research professionals among whom research coordinators and a clinical librarian of the Dutch university medical centres. The percentage of publications that could be assigned to a disease group was compared between WoS research fields. In addition, several institutional profiles resulting from the classification of research output to disease groups were discussed with researchers and deans from those institutions.

Classification of publications by institution and country

The name of an institution is often reported in many different ways in publications. Some authors for example report an abbreviated name while others report the full name, and some authors report the name of the university with which a hospital is associated while other authors report only the name of the hospital itself. These inconsistencies are problematic when analysing the research output of institutions. We addressed this problem by relying on the categorisation of affiliations used in the CWTS Leiden Ranking 2014 [15]. In this way we could compare the research output of the 750 largest universities worldwide (based on number of publications in WoS), of 1099 hospitals, and of 46 public research organisations. Publications from all affiliations, also those not included in the selected institutions, were included when comparing research output between countries.

Publications were assigned fractionally to institutions and countries. This was done based on the number of addresses in the address list of a publication in which a certain institution or country is mentioned. For instance, if a publication includes five addresses and two of these addresses mention Leiden University (e.g., two different departments within Leiden University), the publication is assigned to Leiden University with a weight of 2 / 5 = 0.4. So

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

the publication is not counted as a full publication for Leiden University but as 40% of a full publication. This methodology is known as address-level fractional counting [16].

Indicators of quantity and quality of research

We used several indicators of quantity and quality of biomedical research per disease group to provide quantitative insight in the research strengths of specific institutions and countries. Quantity was measured by the fractionally counted volume of publications of an institution or country. Citations are often seen as an indicator of scientific impact, or somewhat less precisely, as an indicator of quality. Since research fields differ in citation practices, comparison of citation counts between fields is difficult. Likewise, comparison of citation counts between older and more recent publications is problematic, because older publications have had more time to accumulate citations. To overcome this, we identified for each combination of a disease group and a publication year the 10% most cited publications globally. We used the volume of these 'top publications' as an indicator of quality of output when comparing countries or institutions. Only publications that appeared between 2000 and 2012 were used to identify 'top publications', since publications after 2012 were too recent for the calculation of meaningful citation statistics in 2014. Self-citations, that is, citations given by an author to his or her own work, were excluded. For the comparison of research portfolios between countries, between institutions, and over time, we used an institution's (or country's) share in the global publication volume per disease group as an indicator of the total volume (quantity). Additionally, we used the share of top publications in the total output of an institution (or country) as a size-independent indicator for quality. This relative measure enables a comparison of research output for different disease groups within the research portfolio of an institution (or country).

RESULTS

This section first describes the results of the validation of our method. Second, results for several applications of the method are described.

Validation of the proposed method

We were able to relate 54% of all publications in the selected 84 research fields to a disease group, 3.2 million publications in total. Of all publications, 29% were assigned to a single disease group, 14% to two disease groups, and 11% to three or more disease groups. Fields of research with a large share of disease-specific publications were mainly clinical research fields. Over 80% of all publications in research fields such as allergology, rheumatology, and clinical neurology were linked to a disease group. Research fields like ethics, microscopy, and biophysics had a much lower percentage of disease-specific publications (10%, 17%, and 27%, respectively). In these fields, we indeed would not expect a large share of the publications to be linked to a disease group, so the low percentages confirm that our method behaves as expected. We refer to appendix 1 for an overview of the share of disease-specific publications per research field.

Between 2000 and 2012, the annual volume of publications within the included research fields increased by 64%. In the same period, the volume of disease-specific publications increased by 92%. This means that disease-specific publications grew in share: from 48% in 2000 to 57% of total volume in 2012. After manual verification, we found that 2% of the sample of disease-specific publications (n=680) were incorrectly assigned to a disease group, and 1% of the sample of uncategorised publications (n=315) were incorrectly not

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

assigned to a disease group, both indicating the method to be accurate. Incorrect links were mainly due to sentences such as “patients with diabetes were excluded” in the abstracts of publications.

About 1900 institutions were analysed in this study. Together these institutions accounted for 69% of the address lines in disease-specific publications worldwide. 78% of the disease-specific publications had at least one author from one of these institutions.

As expected, we found strong differences in the share of disease-specific publications between different types of research institutions in the Netherlands. We verified institution-specific results with researchers and deans of five top ranking institutions in the Netherlands and abroad. In all cases the disease-specific research output was in line with their expectations about their own institution’s position in relation to other institutions worldwide.

Application 1: Biomedical research output by disease group

Using our method, we can compare the research output between disease groups. The number of publications in the period 2000-2012 varies widely between disease groups, as shown in figure 1. ‘Other infectious diseases (not including HIV and tuberculosis)’ was the disease group with most publications. ‘Diabetes mellitus’, ‘metabolic diseases’, and ‘mood disorders’ were also large. The number of publications on malignant neoplasms was just a little bigger than the total publication volume on heart diseases.

[FIGURE 1]

Interestingly, the worldwide research profile by disease is not constant over time. Some disease groups have seen a rapid growth in research output, while other disease groups have grown only mildly in research output, as shown in figure 2. Lifestyle diseases (obesity and diabetes), cancers (lung, prostate, colon and breast), and mental disorders (depression and other mental disorders) gained in share in the worldwide research portfolio. On the other hand, diseases such as anaemia, pain in chest and throat, leukaemia, and HIV show a decreasing share in the total research portfolio, although the research output has still grown in absolute volume.

[FIGURE 2]

Application 2: Biomedical research output by disease by country

The most cited disease-specific research publications originate from a small set of countries. Figure 3 shows the relative share of countries in the 10% most cited publications per disease group. The top ten countries with the largest share in top 10% most cited research output account for 83% of the total body of disease-specific publications worldwide. Notably, the USA accounts for 45% of the top 10% most cited publications. There are however differences in research profiles between countries. For instance Canada has equal shares in top publications on ‘depression’ and ‘stroke’, while China has twice as many top publications on ‘stroke’ compared to ‘depression’.

[FIGURE 3]

It is possible to evaluate the development over time of each country’s share in publication volume for a specific disease. Figure 4 shows the growth in number of breast cancer publications by country between 2000 and 2012. Although the number of publications of every country has grown during this period, some countries have grown faster than others.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Most western countries have grown slower than the world average. Countries that have grown faster than average are mainly BRIC countries, with China showing 700% growth. Notably, Iran experienced a remarkable 3,500% growth in research output, but its total volume of disease-specific publications remains small.
[FIGURE 4]

Application 3: Research output by disease on an institution level
Our method allows for identification of institutions with a remarkable position in research on a specific disease group. We use Multiple Sclerosis (MS) as an example, but figure 5 can easily be constructed for all 269 disease groups used in this study. The figure shows for all institutions their volume of MS publications and their respective share in the top 10% most cited publications about MS worldwide. Harvard's unique position in MS research is illustrated by the facts that Harvard had the largest share in the total MS publication volume and that one in four of its publications was in the top 10% most cited publications about MS. Other centres with remarkable quantity and quality of MS research were University College London and VU University Amsterdam. A display like figure 5 recognises institutions that have a high quality without a high production.
[FIGURE 5]

Using our method it is possible to follow the research output of individual institutions for specific disease groups over time. As an example, figure 6 shows the rise of South African research output on HIV. Between 2000 and 2004, the annual South African research output on HIV is relatively constant, but from 2005 onward, several South African universities have grown rapidly, passing several famous HIV research institutions in volume. This growth seems partly due to growth of international collaboration. For instance, 10% of all South African publications on HIV were co-authored with Harvard University in 2012, while this was only 2% in 2005. During this time, internationally renowned Harvard scientists such as Bruce Walker and Till Barnighausen have started working part time for the University of Kwazulu-Natal.
[FIGURE 6]

In addition to comparing institutions for a specific disease, our method also allows us to map research portfolios of countries or institutions by disease, based on volume and top 10% publications. Using these portfolio maps, we can now compare complete disease-specific research portfolios between institutions. As an example, we plotted portfolio maps of four universities in figure 7. Substantial differences in their profiles can easily be seen. Harvard University has much larger publication volumes than the three others. Imperial College has a large number of disease groups with at least 30% of their publications counting as top publications. Both University of Amsterdam and Karolinska Institute have a remarkable position in research on malignant oesophageal neoplasms, whereas Imperial College does not.
[FIGURE 7]

DISCUSSION
Our proposed method allows for systematic classification of publications in WoS to disease groups. We were able to classify 54% of all 6.5 million biomedical publications in the WoS database to a disease group. Between 2000 and 2012, 'Other infectious diseases' was the largest disease group with 337,485 publications. In this period, lifestyle diseases, cancers,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

and mental disorders have grown most in research output. On a country level, the USA was responsible for the largest number of top 10% most cited publications per disease group, with a global share of 45%. Iran (+3,500%) and China (+700%) have grown most in research volume. On an institution level, we were able to relate 78% of biomedical publications to a specific institution. Below we describe some examples of potential use and then discuss possibilities for future research.

Potential value of the proposed method

The method can be used for evaluation of historic research performance at the level of specific diseases. It can support decision making in management of research portfolios, showing relative strengths and weaknesses of institutions and countries. Combining these insights with indicators of innovation and research productivity [17] can illustrate whether research performance is aligned with successful transfer of scientific knowledge to clinical practice.

Linking the disease-specific research output to burden of disease provides insights in 'white spots' in global and regional research [18]. These insights can support fact-based allocation of research funding, making it possible to better align research portfolios to local or global needs and to adjust portfolios to changes of those needs over time. This can be the starting point for further understanding of what drives research output other than burden of disease, for instance; economic strengths, political structures, research legacy, etc. Quantitatively unravelling the different drivers that determine disease-specific publication volume could provide insights in how we can realign research efforts across countries to have greater impact on reduction of disease burden.

Opportunities for additional research

Using disease groups based on the ICD-10 classification has the advantage of being exhaustive: all diseases can be included. When looking for research on a rare disease, the used classification system is not specific enough. However, our method can be adapted to answer such specific questions by using specific author keywords and tailor-made text phrases to look for in titles and abstracts. Addition of MeSH descriptors next to author keywords can further complete the method, although this requires the use of other bibliographic databases, since WoS does not include MeSH descriptors. Ultimately, the use of dynamic and customised research categories will make it easier to find the institutions with the strongest positions in research on specific diseases, thus answering portfolio questions in ways that are not possible yet.

Our method classifies each publication to disease nomenclature but does not categorise the nature of disease-specific research. For example, a publication classified to a disease group could describe a new gene involved in the pathogenesis, analyse the societal impact of the disease, or merely state the disease as a potential application for a new surgical technique. Ideally, the method should be supplemented with additional categories that, based on text mining, can identify the type of research and application. Also clinical trial registers (e.g. https://www.clinicaltrialsregister.eu/ or https://clinicaltrials.gov/) can be included. As an example, using a simple algorithm based on MeSH descriptors, it is possible to identify cell-based, animal-based, and patient-based research [19].

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Now that publications are allocated to disease groups, bibliometric indicators of research quantity and quality can be combined with other information available on disease level. For instance, quality of care, patient reported health outcomes, cost of treatment, and patents. This can be valuable in aligning research and health care portfolios of university medical centres.

Conclusion

We have shown that it is possible to systematically link research output to disease groups. Our method makes it possible to compare research output by countries or institutions and to monitor changes in biomedical research output over time or by disease. The novelty and value of the method is that it allows a disease-specific analysis, for instance making it possible to compare research output with burden of disease. Since the major goal of biomedical research is alleviation of disease burden, our method allows for evaluating current strengths and shortcomings.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests disclosed

We have read and understood BMJ policy on declaration of interests and declare no competing interests.

Individual contributions

LL and NH made the definitions of disease groups, categorised the author keywords, and made the disease-specific keywords. TK, NH and LL performed the analysis. LL wrote the manuscript together with NH and TK. NH and LL validated the results with researchers and deans. LW implemented the text mining algorithm, assigned the publications to disease groups, and calculated the bibliometric statistics. The Centre for Science and Technology Studies (CWTS) at Leiden University provided the cleaned address data for the universities, hospitals and public research organisations included in the study. IM, LW and AG provided feedback on the manuscript.

Acknowledgements

The authors would like to thank the research coordinators and deans of the Dutch university medical centres for their contribution to the validation of this research method, the group of medical interns for their assistance in drafting the disease-specific terms, and prof. dr. Marcel Levi for his comments on the method.

Data sharing statement

Technical appendix can be provided. The appendix includes a definition of biomedical research by WoS research fields.

REFERENCE LIST

1. www.ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges/, date accessed: February 2016.

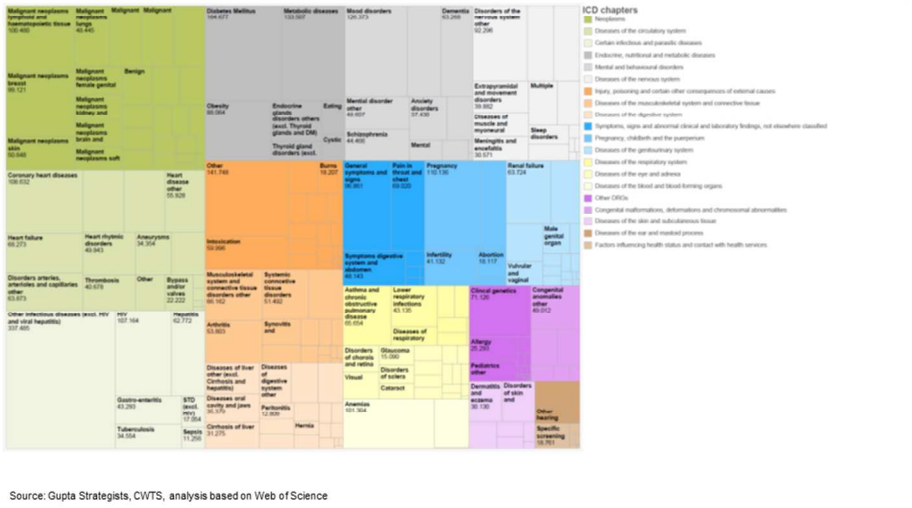
For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

2. Røttingen JA, Regmi S, Eide M, et al. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *Lancet* 2013:382(10):1286-1307.
3. Thomson Reuters. Web of Science. http://www.webofknowledge.com, date accessed: February 2016.
4. Lipscomb C. Medical Subject Headings (MeSH). *Bull Med Libr Assoc*. 2000:88(3):265-266.
5. Evans J, Shim J, Ioannides J. Attention to local health burden and the global disparity of health research. *PLoS ONE* 2012:9(4):e90147.
6. Gillum L, Gouveia C, Dorsey E, et al. NIH Disease funding levels and burden of disease. *PLoS ONE* 2011;6(2):e16837.
7. Gross CP, Anderson GF, Powe NR. The relation between funding by the National Institutes of Health and the burden of disease. *N Engl J Med* 1999:340:1881-1887.
8. Bundschuh M, Groneberg D, Klingelhoefer D, et al. Yellow fever disease: density equalizing mapping and gender analysis of international research output. *Parasites and Vectors* 2013:6:331-43.
9. Adams et al. Thomson Reuters Global Research Report, 2012.
10. Minet Kinge J, Roxrud I, Volsset SE, et al. Are the Norwegian health research investments in line with the disease burden? *Health Res Policy Syst.* 2014:12:64.
11. Lascurain-Sánchez ML, García-Zorita C, Martín-Moreno C, et al. Impact of health science research on the Spanish health system, based on bibliometric and healthcare indicators. *Scientometrics* 2008:77:131
12. King D. The scientific impact of nations. Nature 2004:430:311-316.
13. Moses III H, Matheson D, Cairns-Smith S, et al. The anatomy of Medical Research, US and international comparisons. *JAMA* 2015:313(2);174-189.
14. Rodrigues-Esteban R. Biomedical text mining and its applications. *PLoS Comput Biol.* 2009 Dec;5(12):e1000597.
15. Waltman L, Calero-Medina C, Kosten J, et al. The Leiden ranking 2011/2012: Data collection, indicators and interpretation. *J. Am. Soc. Inf. Sci. Technol.* 2012:63(12):2419-2432.
16. Waltman L, Van Eck N. Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of informetrics* 2015:9(4):872-894.
17. Balas EA and Elkin PL. Technology Transfer from biomedical research to clinical practice: measuring innovation performance. *Eval Health Prof.* 2013;36(4):505-17.
18. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015. http://dx.doi.org/10.1016/S0140-6736(15)60692-4.
19. Weber G. Identifying translational science within the triangle of biomedicine. *J Transl Med* 2013;11:126-36.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 1



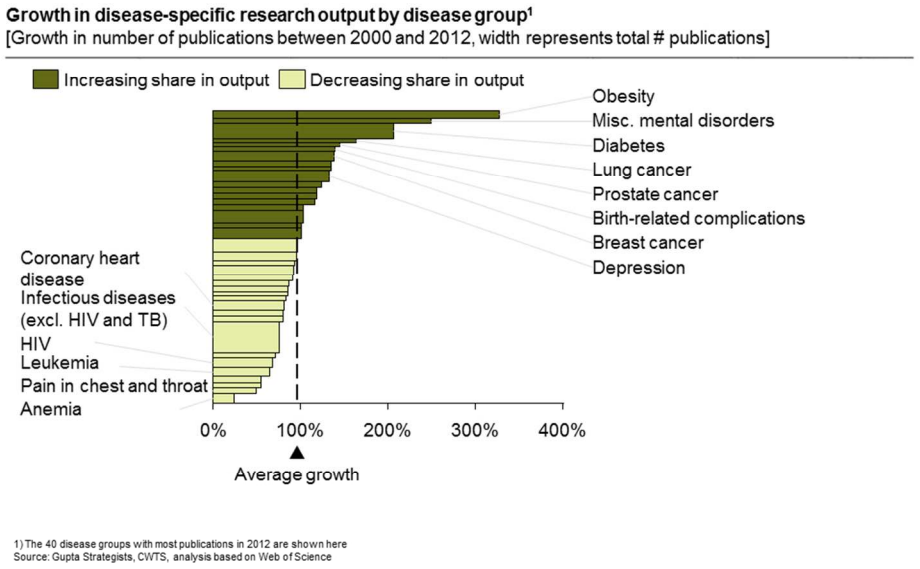
Research output per disease group
[Total volume of publications per disease group between 2000 and 2012]

275x190mm (96 x 96 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 2



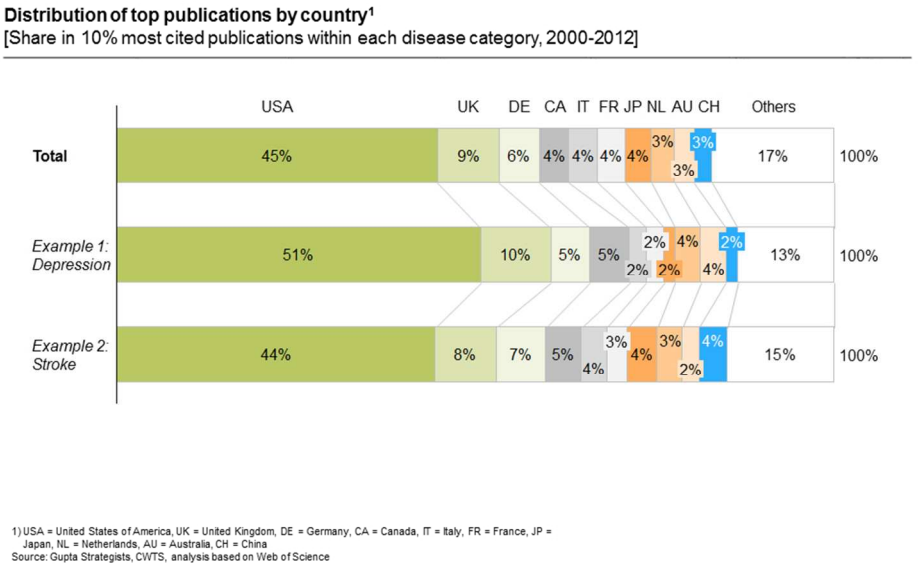
Growth in disease-specific research output by disease group
[Growth in number of publications between 2000 and 2012, width represents total # publications]

275x190mm (96 x 96 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 3



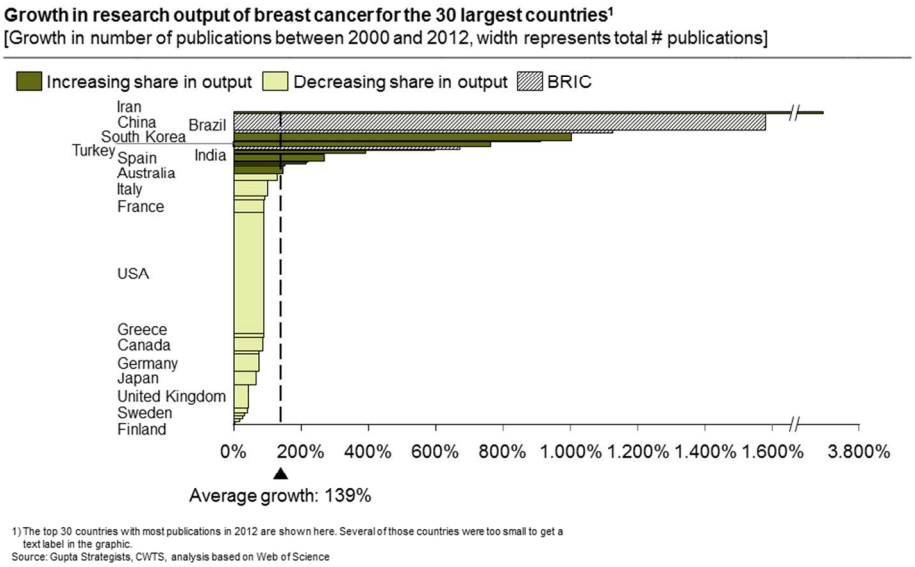
Distribution of top publications by country
[Share in 10% most cited publications within each disease category, 2000-2012]

275x190mm (96 x 96 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 4



Growth in research output of breast cancer for the 30 largest countries
[Growth in number of publications between 2000 and 2012, width represents total # publications]

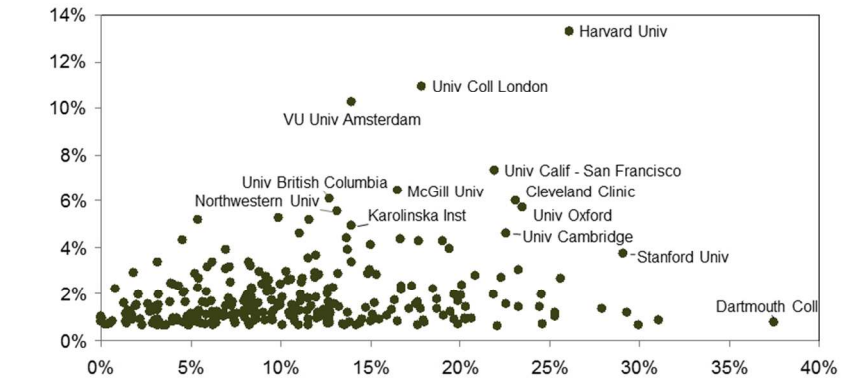
275x190mm (96 x 96 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 5

Scientific output of Multiple Sclerosis by institution[1]
[y-axis: share of institute's output in total output, x-axis: % of publications in 10% most cited, 2000-2012]



1) Institutions with at least 20 publications about MS between 2000 and 2012 are shown
Source: Gupta Strategists, CWTS, analysis based on Web of Science

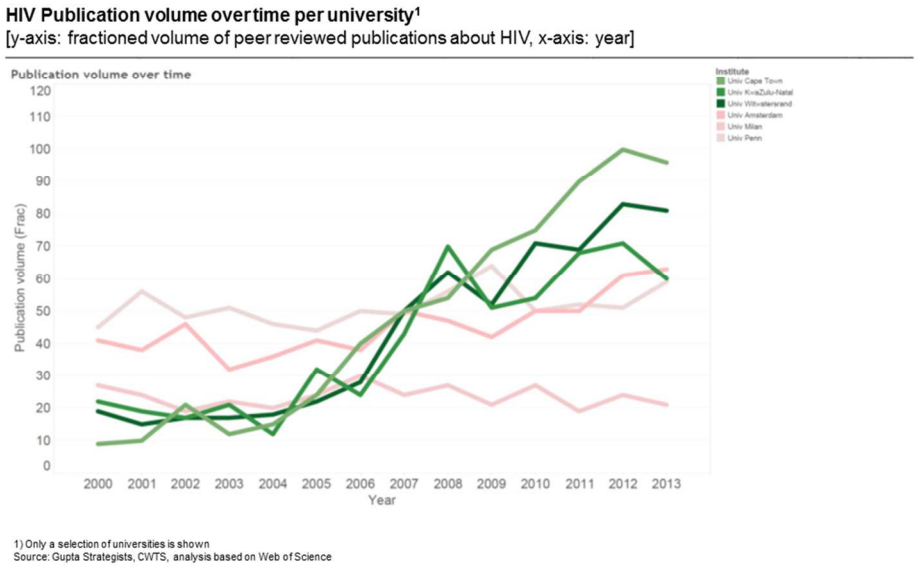
Scientific output of Multiple Sclerosis by institution
[y-axis: share of institute's output in total output, x-axis: % of publications in 10% most cited, 2000-2012]

275x190mm (96 x 96 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 6



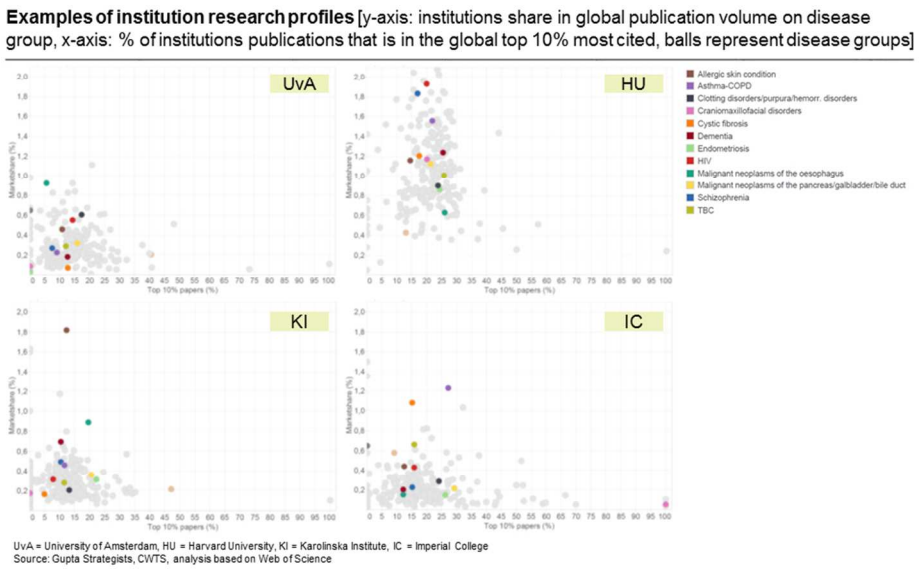
HIV Publication volume over time per university
[y-axis: fractioned volume of peer reviewed publications about HIV, x-axis: year]

275x190mm (96 x 96 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 7



Examples of institution research profiles [y-axis: institutions share in global publication volume on disease group, x-axis: % of institutions publications that is in the global top 10% most cited, balls represent disease groups]

275x190mm (96 x 96 DPI)

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open

IMPROVING THE EVALUATION OF WORLDWIDE BIOMEDICAL RESEARCH OUTPUT: CLASSIFICATION METHOD AND STANDARDIZED BIBLIOMETRIC INDICATORS BY DISEASE

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-020818.R1
Article Type:	Research
Date Submitted by the Author:	16-Apr-2018
Complete List of Authors:	van de Laar, Lissy; Gupta Strategists, de Kruif, Thijs; Gupta Strategists Waltman, Ludo; Leiden University, Centre for Science and Technology Studies Meijer, Ingeborg; Universiteit Leiden, CWTS Gupta, Anshu; Gupta Strategists Hagenaars, Niels; Gupta Strategists
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Health policy, Health informatics, Medical management
Keywords:	HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS, HEALTH ECONOMICS

SCHOLARONE™
Manuscripts

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

IMPROVING THE EVALUATION OF WORLDWIDE BIOMEDICAL RESEARCH OUTPUT: CLASSIFICATION METHOD AND STANDARDIZED BIBLIOMETRIC INDICATORS BY DISEASE

Corresponding author
Lissy van de Laar, MSc
Gupta Strategists, PO Box 16, 4060 GA Ophemert
lissy.vandelaar@gupta.nl
0031 6 34 59 35 07

Co-authors
Ir. Thijs de Kruif, Gupta Strategists, The Netherlands
Dr. Ludo Waltman, Centre for Science and Technology Studies, Leiden University, The Netherlands
Dr. Ingeborg Meijer, Centre for Science and Technology Studies, Leiden University, The Netherlands
Dr. Anshu Gupta, Gupta Strategists, The Netherlands
Dr. Niels Hagenaars, Gupta Strategists, The Netherlands

Key words
Bibliometrics [MeSH], Data mining [MeSH], Classification [MeSH]

Word count excluding title page, abstract, references, figures and tables
3273 words

CONFIDENTIAL
For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

ABSTRACT

Objective: Since most biomedical research focuses on a specific disease, evaluation of research output requires disease-specific bibliometric indicators. Currently used methods are insufficient. The aim of this study is to develop a method that enables detailed analysis of worldwide biomedical research output by disease.

Design: We applied text mining techniques and analysis of author keywords to link publications to disease groups. Fractional counting was used to quantify disease-specific biomedical research output of an institution or country. We calculated global market shares of research output as a relative measure of publication volume. We defined `top publications` as the top 10% most cited publications per disease group worldwide. We used the percentage of publications from an institution or country that were top publications as an indicator of research quality.

Results: We were able to classify 54% of all 6.5 million biomedical publications in our database (based on Web of Science) to a disease group. We could classify 78% of these publications to a specific institution. We show that between 2000 and 2012 'Other infectious diseases' was the largest disease group with 337,485 publications. Lifestyle diseases, cancers, and mental disorders have grown most in research output. The USA was responsible for the largest number of top 10% most cited publications per disease group, with a global share of 45%. Iran (+3,500%) and China (+700%) have grown most in research volume.

Conclusions: The proposed method provides a tool to assess biomedical research output in new ways. It can be used for evaluation of historic research performance, to support decision making in management of research portfolios, and to allocate research funding. Furthermore, using this method to link disease-specific research output to burden of disease can contribute to a better understanding of the societal impact of biomedical research.

STRENGTHS AND LIMITATIONS OF THIS STUDY

Strengths

- The proposed method offers quantitative insight in research quantity and quality for 269 disease groups.
- The proposed method can be used for evaluation of historic research performance at disease level. It can support decision making in management of research portfolios, showing relative strengths and weaknesses of institutions and countries, as well as identifying research gaps at national and global level. It can also be valuable in allocation of research funding.

Limitations

- Author keywords were used instead of the standardised MeSH descriptors, which are not available in the Web of Science database.
- Research about for instance molecular mechanisms, medical techniques, and health sciences could often not be classified to a specific disease group and was thus not included in our results.

CONFIDENTIAL
For peer review only

INTRODUCTION

One of the goals of biomedical research is to eradicate burden of disease. The grand societal challenges in European funding also build on the premise that (biomedical) research should contribute to prevention and treatment of diseases [1].

Yet surprisingly, biomedical research output has not been systematically catalogued by diseases so far [2]. Most publicly available metrics for analysing biomedical research by topic have severe limitations. Research fields in the Web of Science database produced by Clarivate Analytics are defined at a too high level, since they cover a complete medical specialism [3]. The Scopus database produced by Elsevier has the same problem. Medical Subject Headings (MeSH) terms [4] are more specific, but are available only for a selection of journals.

Several authors have made efforts to analyse research output and funding at disease level, but only for a selection of diseases. Evans et al compared research output between countries for 19 disease groups, based on the International Classification of Disease (ICD)-9 chapters [5]. Gillum et al [6] and Gross et al [7] analysed burden of disease and research funding for a selection of 29 conditions, derived from the ICD. In various other studies funding, research output and burden of disease were described for specific diseases in a case by case approach. This was done for example for yellow fever [8] and neglected tropical diseases [9]. In other studies, total biomedical research output was analysed for specific countries [10, 11] or compared between countries [12, 13].

Text mining techniques are increasingly applied to biomedical text to uncover unseen relationships [14]. In this study we use these techniques to create a reference structure of disease groups and to catalogue publications accordingly. This opens a bridge between biomedical research output and other information available at disease level, which can contribute to a better understanding of the societal impact of biomedical science.

METHODS

Selection of biomedical publications

The analysis was based on the Clarivate Analytics WoS database available at the Centre for Science and Technology Studies (CWTS) of Leiden University. Since the goal of this study is to quantify research output by disease, we included biomedical research fields only. Of the 250 WoS research fields, we selected the 84 fields that are most medically oriented. We validated the selection by looking at the research output of the eight Dutch university medical centres: over 98% of their publications were in one of these fields. Appendix 1 provides a full list of research fields included in this study. The dataset was compiled in June 2014. It includes all publications in the 84 selected research fields, published between 2000 and early 2014, with WoS document type ‘article’ or ‘review’. Not all publications from the first six months of 2014 were available, due to periodical updating of the CWTS in-house version of the WoS database. The dataset contained 6.5 million publications in total.

Classification of publications by disease group

We defined 269 disease groups, based on the ICD-10 classification and covering the full spectrum of this classification. We used a two-step approach to categorise publications to disease groups.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

For peer review only CONFIDENTIAL

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

First, we categorised the author keywords listed by authors in their publications. In total, 158,700 unique author keywords were used in at least ten publications in our dataset. Of these keywords, the 32,400 most frequently used keywords (used in more than 70 publications each) were short listed and further evaluated. 21% of these keywords were specific for a single disease group. For example, the keyword 'Alzheimer's disease' was linked to 'dementia'. Many keywords were not suitable to use for categorisation to disease groups because they were either too general or not disease-specific. Examples of keywords not linked to a disease group are 'inflammation' and 'keyhole surgery'. We note that not all publications include author keywords.

In the second step, a text mining algorithm was used to search for disease-specific terms in titles and abstracts of publications. In this step, first a list of 10,983 unambiguous, disease-specific terms was generated by hand by medical professionals to characterize specific disease groups. Examples of terms for the disease group 'malignant neoplasm prostate' include 'prostate cancer', 'prostate carcinoma', 'malignant tumor prostate', and 'sarcoma prostate'. The generated disease-specific terms were then reviewed by another medical professional for ambiguity. Subsequently publications with one of these 10,983 terms in either title or abstract were assigned to the corresponding disease group. If the same publication was assigned to multiple disease groups, it was fully counted for all of them.

The method was validated in several ways. The first step was a manual examination of a random sample of 680 publications assigned to a disease group. Subsequently, a random sample of 315 publications not assigned to a disease group was manually examined. The examination was executed by research professionals among whom research coordinators and a clinical librarian of the Dutch university medical centres. The percentage of publications that could be assigned to a disease group was compared between WoS research fields. In addition, several institutional profiles resulting from the classification of research output to disease groups were discussed with researchers and deans from those institutions.

Classification of publications by institution and country

The name of an institution is often reported in many different ways in publications. Some authors for example report an abbreviated name while others report the full name, and some authors report the name of the university with which a hospital is associated while other authors report only the name of the hospital itself. These inconsistencies are problematic when analysing the research output of institutions. We addressed this problem by relying on the categorisation of affiliations used in the CWTS Leiden Ranking 2014 [15]. In this way we could compare the research output of the 750 largest universities worldwide (based on number of publications in WoS), of 1099 hospitals, and of 46 public research organisations. Publications from all affiliations, also those not included in the selected institutions, were included when comparing research output between countries.

Publications were assigned fractionally to institutions and countries. This was done based on the number of addresses in the address list of a publication in which a certain institution or country is mentioned. For instance, if a publication includes five addresses and two of these addresses mention Leiden University (e.g., two different departments within Leiden University), the publication is assigned to Leiden University with a weight of 2 / 5 = 0.4. So

the publication is not counted as a full publication for Leiden University but as 40% of a full publication. This methodology is known as address-level fractional counting [16].

Indicators of quantity and quality of research

We used several indicators of quantity and quality of biomedical research per disease group to provide quantitative insight in the research strengths of specific institutions and countries. Quantity was measured by the fractionally counted volume of publications of an institution or country. Citations are often seen as an indicator of scientific impact, or somewhat less precisely, as an indicator of quality. Since research fields differ in citation practices, comparison of citation counts between fields is difficult. Likewise, comparison of citation counts between older and more recent publications is problematic, because older publications have had more time to accumulate citations. To overcome this, we identified for each combination of a disease group and a publication year the 10% most cited publications globally. We used the volume of these 'top publications' as an indicator of quality of output when comparing countries or institutions. Only publications that appeared between 2000 and 2012 were used to identify 'top publications', since publications after 2012 were too recent for the calculation of meaningful citation statistics in 2014. Self-citations, that is, citations given by an author to his or her own work, were excluded. For the comparison of research portfolios between countries, between institutions, and over time, we used an institution's (or country's) share in the global publication volume per disease group as an indicator of the total volume (quantity). Additionally, we used the share of top publications in the total output of an institution (or country) as a size-independent indicator for quality. This relative measure enables a comparison of research output for different disease groups within the research portfolio of an institution (or country).

Patient and Public Involvement

No patients or public were involved in our study.

RESULTS

This section first describes the results of the validation of our method. Second, results for several applications of the method are described.

Validation of the proposed method

We were able to relate 54% of all publications in the selected 84 research fields to a disease group, 3.2 million publications in total. Of all publications, 29% were assigned to a single disease group, 14% to two disease groups, and 11% to three or more disease groups. Fields of research with a large share of disease-specific publications were mainly clinical research fields. Over 80% of all publications in research fields such as allergology, rheumatology, and clinical neurology were linked to a disease group. Research fields like ethics, microscopy, and biophysics had a much lower percentage of disease-specific publications (10%, 17%, and 27%, respectively). In these fields, we indeed would not expect a large share of the publications to be linked to a disease group, so the low percentages confirm that our method behaves as expected. We refer to appendix 1 for an overview of the share of disease-specific publications per research field.

Between 2000 and 2012, the annual volume of publications within the included research fields increased by 64%. In the same period, the volume of disease-specific publications increased by 92%. This means that disease-specific publications grew in share: from 48% in

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

CONFIDENTIAL For peer review only

2000 to 57% of total volume in 2012. After manual verification, we found that 2% of the sample of disease-specific publications (n=680) were incorrectly assigned to a disease group, and 1% of the sample of uncategorised publications (n=315) were incorrectly not assigned to a disease group, both indicating the method to be accurate. Incorrect links were mainly due to sentences such as “patients with diabetes were excluded” in the abstracts of publications.

About 1900 institutions were analysed in this study. Together these institutions accounted for 69% of the address lines in disease-specific publications worldwide. 78% of the disease-specific publications had at least one author from one of these institutions.

As expected, we found strong differences in the share of disease-specific publications between different types of research institutions in the Netherlands. We verified institution-specific results with researchers and deans of five top ranking institutions in the Netherlands and abroad. In all cases the disease-specific research output was in line with their expectations about their own institution’s position in relation to other institutions worldwide.

Application 1: Biomedical research output by disease group

Using our method, we can compare the research output between disease groups. The number of publications in the period 2000-2012 varies widely between disease groups, as shown in figure 1. ‘Other infectious diseases (not including HIV and tuberculosis)’ was the disease group with most publications. ‘Diabetes mellitus’, ‘metabolic diseases’, and ‘mood disorders’ were also large. The number of publications on malignant neoplasms was just a little bigger than the total publication volume on heart diseases.

[FIGURE 1]

Interestingly, the worldwide research profile by disease is not constant over time. Some disease groups have seen a rapid growth in research output, while other disease groups have grown only mildly in research output, as shown in figure 2. Lifestyle diseases (obesity and diabetes), cancers (lung, prostate, colon and breast), and mental disorders (depression and other mental disorders) gained in share in the worldwide research portfolio. On the other hand, diseases such as anaemia, pain in chest and throat, leukaemia, and HIV show a decreasing share in the total research portfolio, although the research output has still grown in absolute volume.

[FIGURE 2]

Application 2: Biomedical research output by disease by country

The most cited disease-specific research publications originate from a small set of countries. Figure 3 shows the relative share of countries in the 10% most cited publications per disease group. The top ten countries with the largest share in top 10% most cited research output account for 83% of the total body of disease-specific publications worldwide. Notably, the USA accounts for 45% of the top 10% most cited publications. There are however differences in research profiles between countries. For instance Canada has equal shares in top publications on ‘depression’ and ‘stroke’, while China has twice as many top publications on ‘stroke’ compared to ‘depression’.

[FIGURE 3]

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

It is possible to evaluate the development over time of each country's share in publication volume for a specific disease. Figure 4 shows the growth in number of breast cancer publications by country between 2000 and 2012. Although the number of publications of every country has grown during this period, some countries have grown faster than others. Most western countries have grown slower than the world average. Countries that have grown faster than average are mainly BRIC countries, with China showing 700% growth. Notably, Iran experienced a remarkable 3,500% growth in research output, but its total volume of disease-specific publications remains small.
[FIGURE 4]

Application 3: Research output by disease on an institution level

Our method allows for identification of institutions with a remarkable position in research on a specific disease group. We use Multiple Sclerosis (MS) as an example, but figure 5 can easily be constructed for all 269 disease groups used in this study. The figure shows for all institutions their volume of MS publications and their respective share in the top 10% most cited publications about MS worldwide. Harvard's unique position in MS research is illustrated by the facts that Harvard had the largest share in the total MS publication volume and that one in four of its publications was in the top 10% most cited publications about MS. Other centres with remarkable quantity and quality of MS research were University College London and VU University Amsterdam. A display like figure 5 recognises institutions that have a high quality without a high production.
[FIGURE 5]

Using our method it is possible to follow the research output of individual institutions for specific disease groups over time. As an example, figure 6 shows the rise of South African research output on HIV. Between 2000 and 2004, the annual South African research output on HIV is relatively constant, but from 2005 onward, several South African universities have grown rapidly, passing several famous HIV research institutions in volume. This growth seems partly due to growth of international collaboration. For instance, 10% of all South African publications on HIV were co-authored with Harvard University in 2012, while this was only 2% in 2005. During this time, internationally renowned Harvard scientists such as Bruce Walker and Till Barnighausen have started working part time for the University of Kwazulu-Natal.
[FIGURE 6]

In addition to comparing institutions for a specific disease, our method also allows us to map research portfolios of countries or institutions by disease, based on volume and top 10% publications. Using these portfolio maps, we can now compare complete disease-specific research portfolios between institutions. As an example, we plotted portfolio maps of four universities in figure 7. Substantial differences in their profiles can easily be seen. Harvard University has much larger publication volumes than the three others. Imperial College has a large number of disease groups with at least 30% of their publications counting as top publications. Both University of Amsterdam and Karolinska Institute have a remarkable position in research on malignant oesophageal neoplasms, whereas Imperial College does not.
[FIGURE 7]

DISCUSSION

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Our proposed method allows for systematic classification of publications in WoS to disease groups. We were able to classify 54% of all 6.5 million biomedical publications in the WoS database to a disease group. Between 2000 and 2012, 'Other infectious diseases' was the largest disease group with 337,485 publications. In this period, lifestyle diseases, cancers, and mental disorders have grown most in research output. On a country level, the USA was responsible for the largest number of top 10% most cited publications per disease group, with a global share of 45%. Iran (+3,500%) and China (+700%) have grown most in research volume. On an institution level, we were able to relate 78% of biomedical publications to a specific institution. Below we describe some examples of potential use and then discuss possibilities for future research.

Potential value of the proposed method

The method can be used for evaluation of historic research performance at the level of specific diseases. It can support decision making in management of research portfolios, showing relative strengths and weaknesses of institutions and countries. Combining these insights with indicators of innovation and research productivity [17] can illustrate whether research performance is aligned with successful transfer of scientific knowledge to clinical practice.

Linking the disease-specific research output to burden of disease provides insights in 'white spots' in global and regional research [18]. These insights can support fact-based allocation of research funding, making it possible to better align research portfolios to local or global needs and to adjust portfolios to changes of those needs over time. This can be the starting point for further understanding of what drives research output other than burden of disease, for instance; economic strengths, political structures, research legacy, etc. Quantitatively unravelling the different drivers that determine disease-specific publication volume could provide insights in how we can realign research efforts across countries to have greater impact on reduction of disease burden.

Opportunities for additional research

Using disease groups based on the ICD-10 classification has the advantage of being exhaustive: all diseases can be included. When looking for research on a rare disease, the used classification system is not specific enough. However, our method can be adapted to answer such specific questions by using specific author keywords and tailor-made text phrases to look for in titles and abstracts. Addition of MeSH descriptors next to author keywords can further complete the method, although this requires the use of other bibliographic databases, since WoS does not include MeSH descriptors. Ultimately, the use of dynamic and customised research categories will make it easier to find the institutions with the strongest positions in research on specific diseases, thus answering portfolio questions in ways that are not possible yet.

Our method classifies each publication to disease nomenclature but does not categorise the nature of disease-specific research. For example, a publication classified to a disease group could describe a new gene involved in the pathogenesis, analyse the societal impact of the disease, or merely state the disease as a potential application for a new surgical technique. Ideally, the method should be supplemented with additional categories that, based on text mining, can identify the type of research and application. Also clinical trial registers (e.g. https://www.clinicaltrialsregister.eu/ or https://clinicaltrials.gov/) can be included. As an

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

example, using a simple algorithm based on MeSH descriptors, it is possible to identify cell-based, animal-based, and patient-based research [19].

Now that publications are allocated to disease groups, bibliometric indicators of research quantity and quality can be combined with other information available on disease level. For instance, quality of care, patient reported health outcomes, cost of treatment, and patents. This can be valuable in aligning research and health care portfolios of university medical centres.

Conclusion

We have shown that it is possible to systematically link research output to disease groups. Our method makes it possible to compare research output by countries or institutions and to monitor changes in biomedical research output over time or by disease. The novelty and value of the method is that it allows a disease-specific analysis, for instance making it possible to compare research output with burden of disease. Since the major goal of biomedical research is alleviation of disease burden, our method allows for evaluating current strengths and shortcomings.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests disclosed

We have read and understood BMJ policy on declaration of interests and declare no competing interests.

Individual contributions

LL and NH made the definitions of disease groups, categorised the author keywords, and made the disease-specific keywords. TK, NH and LL performed the analysis. LL wrote the manuscript together with NH and TK. NH and LL validated the results with researchers and deans. LW implemented the text mining algorithm, assigned the publications to disease groups, and calculated the bibliometric statistics. The Centre for Science and Technology Studies (CWTS) at Leiden University provided the cleaned address data for the universities, hospitals and public research organisations included in the study. IM, LW and AG provided feedback on the manuscript.

Acknowledgements

The authors would like to thank the research coordinators and deans of the Dutch university medical centres for their contribution to the validation of this research method, the group of medical interns for their assistance in drafting the disease-specific terms, and prof. dr. Marcel Levi for his comments on the method.

Data sharing statement

Technical appendix can be provided. The appendix includes a definition of biomedical research by WoS research fields.

REFERENCE LIST

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

1. www.ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges/, date accessed: February 2016.
2. Røttingen JA, Regmi S, Eide M, et al. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *Lancet* 2013:382(10):1286-1307.
3. Thomson Reuters. Web of Science. http://www.webofknowledge.com, date accessed: February 2016.
4. Lipscomb C. Medical Subject Headings (MeSH). *Bull Med Libr Assoc.* 2000:88(3):265-266.
5. Evans J, Shim J, Ioannides J. Attention to local health burden and the global disparity of health research. *PLoS ONE* 2012:9(4):e90147.
6. Gillum L, Gouveia C, Dorsey E, et al. NIH Disease funding levels and burden of disease. *PLoS ONE* 2011;6(2):e16837.
7. Gross CP, Anderson GF, Powe NR. The relation between funding by the National Institutes of Health and the burden of disease. *N Engl J Med* 1999:340:1881-1887.
8. Bundschuh M, Groneberg D, Klingelhoefer D, et al. Yellow fever disease: density equalizing mapping and gender analysis of international research output. *Parasites and Vectors* 2013:6:331-43.
9. Adams et al. Thomson Reuters Global Research Report, 2012.
10. Minet Kinge J, Roxrud I, Volsset SE, et al. Are the Norwegian health research investments in line with the disease burden? *Health Res Policy Syst.* 2014:12:64.
11. Lascurain-Sánchez ML, García-Zorita C, Martín-Moreno C, et al. Impact of health science research on the Spanish health system, based on bibliometric and healthcare indicators. *Scientometrics* 2008:77:131
12. King D. The scientific impact of nations. Nature 2004:430:311-316.
13. Moses III H, Matheson D, Cairns-Smith S, et al. The anatomy of Medical Research, US and international comparisons. *JAMA* 2015:313(2);174-189.
14. Rodrigues-Esteban R. Biomedical text mining and its applications. *PLoS Comput Biol.* 2009 Dec;5(12):e1000597.
15. Waltman L, Calero-Medina C, Kosten J, et al. The Leiden ranking 2011/2012: Data collection, indicators and interpretation. *J. Am. Soc. Inf. Sci. Technol.* 2012:63(12):2419-2432.
16. Waltman L, Van Eck N. Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of informetrics* 2015:9(4):872-894.
17. Balas EA and Elkin PL. Technology Transfer from biomedical research to clinical practice: measuring innovation performance. *Eval Health Prof.* 2013;36(4):505-17.
18. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015. http://dx.doi.org/10.1016/S0140-6736(15)60692-4.
19. Weber G. Identifying translational science within the triangle of biomedicine. *J Transl Med* 2013;11:126-36.

FIGURE LEGENDS

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 1: Research output per disease group. [Total volume of publications per disease group between 2000 and 2012]

Figure 2: Growth in disease-specific research output by disease group. [Growth in number of publications between 2000 and 2012, width represents total # publications. Only the 40 disease groups with the most publications in 2012 are shown.]

Figure 3: Distribution of top publications by country. [Share in 10% most cited publications within each disease category, 2000-2012]

Figure 4: Growth in research output of breast cancer for the 30 largest countries. [Growth in number of publications between 2000 and 2012, width represents total # publications]

Figure 5: Scientific output of Multiple Sclerosis by institution. [y-axis: share of institute`s output in total output, x-axis: % of publications in 10% most cited, 2000-2012. Only institutions with at least 20 publications about MS in study period were shown.]

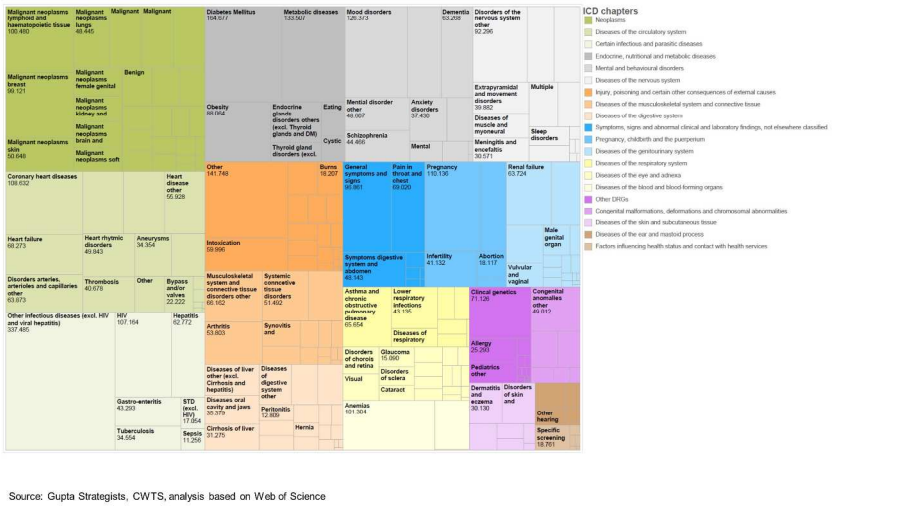
Figure 6: HIV Publication volume over time per university for selected universities. [y-axis: fractioned volume of peer reviewed publications about HIV, x-axis: year]

Figure 7: Examples of institution research profiles. [y-axis: institutions share in global publication volume on disease group, x-axis: % of institutions publications that is in the global top 10% most cited, balls represent disease groups]

CONFIDENTIAL For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 1



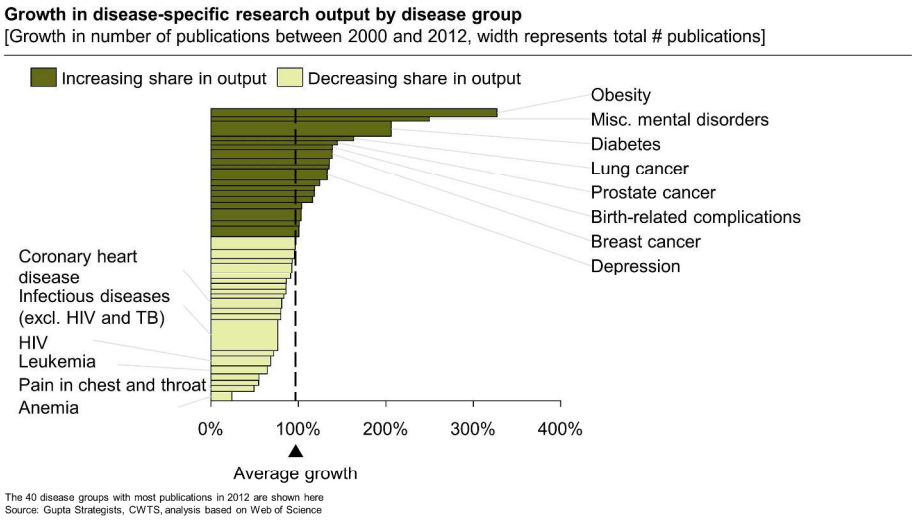
Research output per disease group
[Total volume of publications per disease group between 2000 and 2012]

275x190mm (300 x 300 DPI)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 2



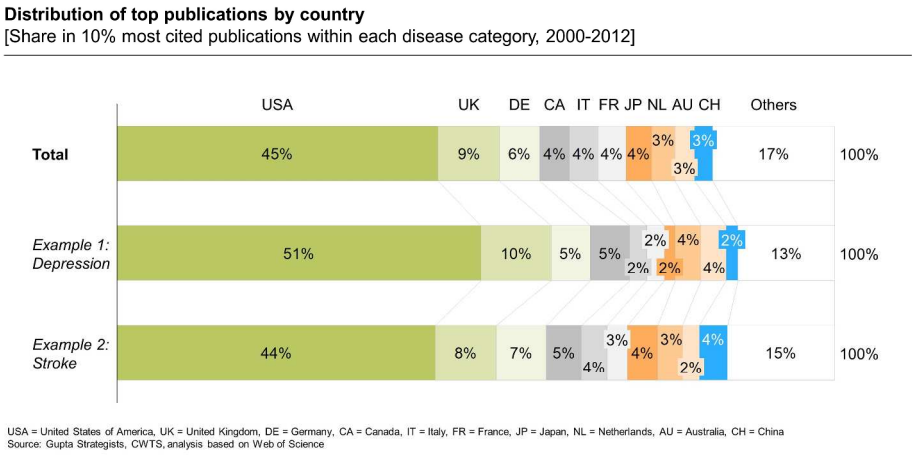
Growth in disease-specific research output by disease group
[Growth in number of publications between 2000 and 2012, width represents total # publications]

275x190mm (300 x 300 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 3



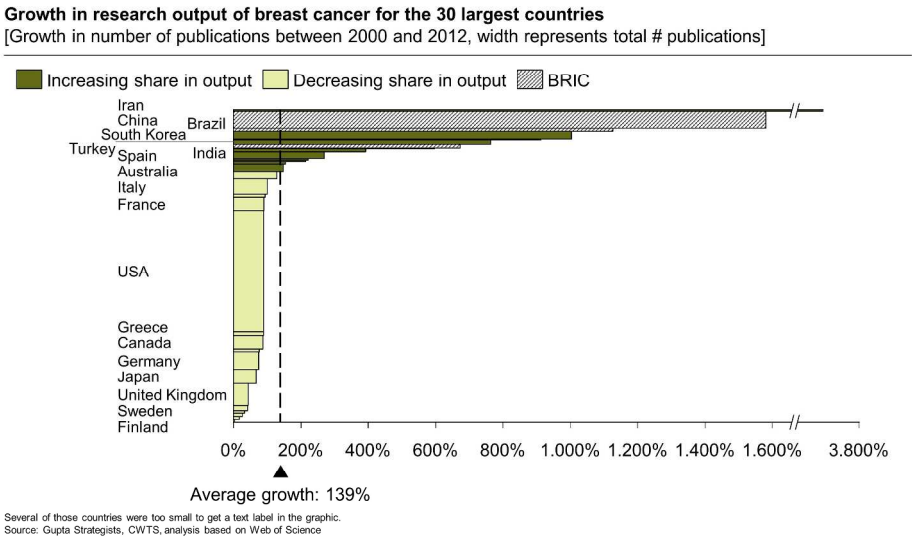
Distribution of top publications by country
[Share in 10% most cited publications within each disease category, 2000-2012]

275x190mm (300 x 300 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 4



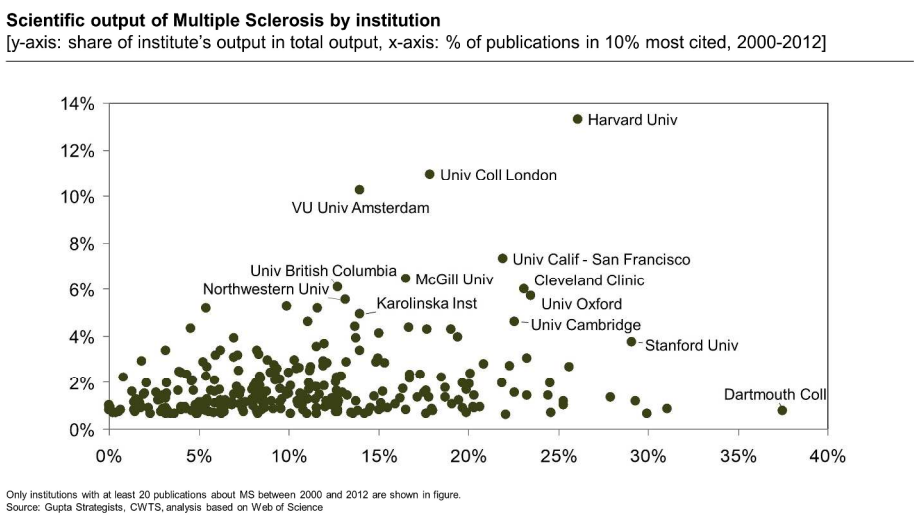
Growth in research output of breast cancer for the 30 largest countries
[Growth in number of publications between 2000 and 2012, width represents total # publications]

275x190mm (300 x 300 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 5



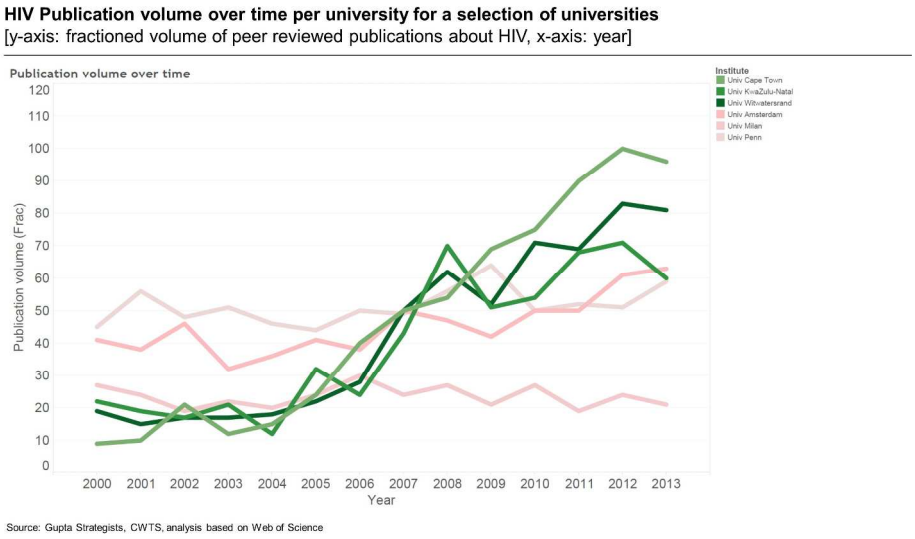
Scientific output of Multiple Sclerosis by institution
[y-axis: share of institute's output in total output, x-axis: % of publications in 10% most cited, 2000-2012]

275x190mm (300 x 300 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 6



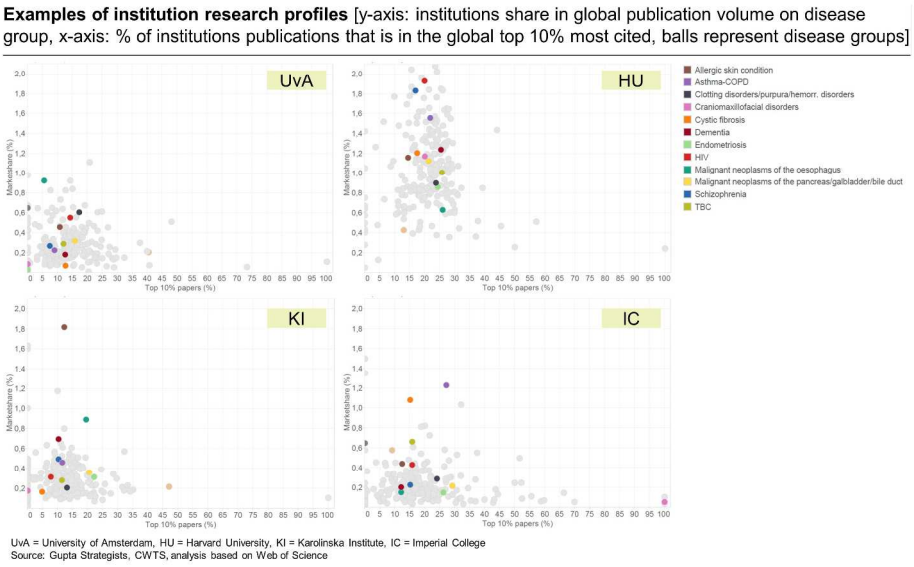
HIV Publication volume over time per university for a selection of universities
[y-axis: fractioned volume of peer reviewed publications about HIV, x-axis: year]

275x190mm (300 x 300 DPI)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 7



Examples of institution research profiles [y-axis: institutions share in global publication volume on disease group, x-axis: % of institutions publications that is in the global top 10% most cited, balls represent disease groups]

275x190mm (300 x 300 DPI)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

APPENDIX 1: Research fields included in our analysis

WoS research field	Total number of publications in study period	% of publications assigned to a diagnosis
ALLERGY	16183	88%
RHEUMATOLOGY	49182	87%
TROPICAL MEDICINE	15421	85%
OBSTETRICS & GYNECOLOGY	88851	83%
INFECTIOUS DISEASES	70945	82%
CARDIAC & CARDIOVASCULAR SYSTEMS	152218	81%
CLINICAL NEUROLOGY	156405	81%
UROLOGY & NEPHROLOGY	113375	80%
GASTROENTEROLOGY & HEPATOLOGY	112952	79%
VIROLOGY	58123	78%
RESPIRATORY SYSTEM	56867	78%
ONCOLOGY	261038	77%
PERIPHERAL VASCULAR DISEASE	77053	76%
PATHOLOGY	61842	76%
PSYCHIATRY	103533	75%
DERMATOLOGY	65094	75%
PEDIATRICS	109175	75%
OPHTHALMOLOGY	87449	74%
OTORHINOLARYNGOLOGY	46286	73%
TRANSPLANTATION	25873	72%
ENDOCRINOLOGY & METABOLISM	144658	71%
MEDICINE, GENERAL & INTERNAL	208138	71%
SURGERY	238955	70%
CRITICAL CARE MEDICINE	32027	69%
HEMATOLOGY	87966	69%

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

REPRODUCTIVE BIOLOGY	28696	67%
ANDROLOGY	4659	67%
ORTHOPEDICS	61249	65%
MEDICINE, RESEARCH & EXPERIMENTAL	106955	64%
IMMUNOLOGY	152354	64%
EMERGENCY MEDICINE	23264	64%
MEDICAL LABORATORY TECHNOLOGY	27827	64%
PARASITOLOGY	34936	62%
GERIATRICS & GERONTOLOGY	26481	60%
NUTRITION & DIETETICS	69840	55%
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH	149568	55%
SUBSTANCE ABUSE	20292	55%
PSYCHOLOGY, CLINICAL	42142	55%
PRIMARY HEALTH CARE	7291	55%
RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING	139186	54%
REHABILITATION	37983	52%
PHARMACOLOGY & PHARMACY	262881	51%
ANESTHESIOLOGY	40227	50%
NEUROSCIENCES	260468	50%
SOCIAL SCIENCES, BIOMEDICAL	10847	
TOXICOLOGY	63389	46%
NEUROIMAGING	9306	45%
DENTISTRY/ORAL SURGERY & MEDICINE	79848	45%
MICROBIOLOGY	146992	42%
GERONTOLOGY	14255	42%
MEDICINE, LEGAL	13686	41%
GENETICS & HEREDITY	133550	41%

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

PHYSIOLOGY	73414	41%
EDUCATION, SPECIAL	6912	40%
NURSING	46557	39%
PSYCHOLOGY, DEVELOPMENTAL	28239	39%
HEALTH CARE SCIENCES & SERVICES	38599	38%
SPORT SCIENCES	52785	37%
HEALTH POLICY & SERVICES	21533	36%
ENGINEERING, BIOMEDICAL	50234	36%
BEHAVIORAL SCIENCES	25766	36%
CELL BIOLOGY	170382	36%
ANATOMY & MORPHOLOGY	15014	34%
CELL & TISSUE ENGINEERING	6192	32%
PSYCHOLOGY, MULTIDISCIPLINARY	94172	31%
MULTIDISCIPLINARY SCIENCES	232494	31%
AUDIOLOGY & SPEECH-LANGUAGE PATHOLOGY	10072	31%
DEVELOPMENTAL BIOLOGY	33875	31%
PSYCHOLOGY, BIOLOGICAL	4084	30%
MEDICAL INFORMATICS	9416	30%
BIOCHEMISTRY & MOLECULAR BIOLOGY	408812	29%
BIOPHYSICS	75287	27%
MEDICAL ETHICS	2301	27%
MATERIALS SCIENCE, BIOMATERIALS	24478	24%
BIOCHEMICAL RESEARCH METHODS	79569	21%
PSYCHOLOGY, PSYCHOANALYSIS	5662	20%
ACOUSTICS	25576	18%
MATHEMATICAL & COMPUTATIONAL BIOLOGY	22087	18%
MICROSCOPY	10412	17%

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

SOCIAL SCIENCES, INTERDISCIPLINARY	24327	14%
PSYCHOLOGY, EXPERIMENTAL	33736	13%
SOCIAL ISSUES	9151	12%
ETHICS	8254	10%
EDUCATION, SCIENTIFIC DISCIPLINES	17290	8%

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

IMPROVING THE EVALUATION OF WORLDWIDE BIOMEDICAL RESEARCH OUTPUT: CLASSIFICATION METHOD AND STANDARDIZED BIBLIOMETRIC INDICATORS BY DISEASE

Corresponding author
Lissy van de Laar, MSc
Gupta Strategists, PO Box 16, 4060 GA Ophemert
lissy.vandelaar@gupta.nl
0031 6 34 59 35 07

Co-authors
Ir. Thijs de Kruif, Gupta Strategists, The Netherlands
Dr. Ludo Waltman, Centre for Science and Technology Studies, Leiden University, The Netherlands
Dr. Ingeborg Meijer, Centre for Science and Technology Studies, Leiden University, The Netherlands
Dr. Anshu Gupta, Gupta Strategists, The Netherlands
Dr. Niels Hagenaars, Gupta Strategists, The Netherlands

Key words
Bibliometrics [MeSH], Data mining [MeSH], Classification [MeSH]

Word count excluding title page, abstract, references, figures and tables
3273 words

CONFIDENTIAL
For peer review only

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

ABSTRACT

Objective: Since most biomedical research focuses on a specific disease, evaluation of research output requires disease-specific bibliometric indicators. Currently used methods are insufficient. The aim of this study is to develop a method that enables detailed analysis of worldwide biomedical research output by disease.

Design: We applied text mining techniques and analysis of author keywords to link publications to disease groups. Fractional counting was used to quantify disease-specific biomedical research output of an institution or country. We calculated global market shares of research output as a relative measure of publication volume. We defined `top publications` as the top 10% most cited publications per disease group worldwide. We used the percentage of publications from an institution or country that were top publications as an indicator of research quality.

Results: We were able to classify 54% of all 6.5 million biomedical publications in our database (based on Web of Science) to a disease group. We could classify 78% of these publications to a specific institution. We show that between 2000 and 2012 'Other infectious diseases' was the largest disease group with 337,485 publications. Lifestyle diseases, cancers, and mental disorders have grown most in research output. The USA was responsible for the largest number of top 10% most cited publications per disease group, with a global share of 45%. Iran (+3,500%) and China (+700%) have grown most in research volume.

Conclusions: The proposed method provides a tool to assess biomedical research output in new ways. It can be used for evaluation of historic research performance, to support decision making in management of research portfolios, and to allocate research funding. Furthermore, using this method to link disease-specific research output to burden of disease can contribute to a better understanding of the societal impact of biomedical research.

STRENGTHS AND LIMITATIONS OF THIS STUDY

Strengths

- The proposed method offers quantitative insight in research quantity and quality for 269 disease groups.
- The proposed method can be used for evaluation of historic research performance at disease level. It can support decision making in management of research portfolios, showing relative strengths and weaknesses of institutions and countries, as well as identifying research gaps at national and global level. It can also be valuable in allocation of research funding.

Limitations

- Author keywords were used instead of the standardised MeSH descriptors, which are not available in the Web of Science database.
- Research about for instance molecular mechanisms, medical techniques, and health sciences could often not be classified to a specific disease group and was thus not included in our results.

CONFIDENTIAL
For peer review only

INTRODUCTION

One of the goals of biomedical research is to eradicate burden of disease. The grand societal challenges in European funding also build on the premise that (biomedical) research should contribute to prevention and treatment of diseases [1].

Yet surprisingly, biomedical research output has not been systematically catalogued by diseases so far [2]. Most publicly available metrics for analysing biomedical research by topic have severe limitations. Research fields in the Web of Science database produced by Clarivate Analytics are defined at a too high level, since they cover a complete medical specialism [3]. The Scopus database produced by Elsevier has the same problem. Medical Subject Headings (MeSH) terms [4] are more specific, but are available only for a selection of journals.

Several authors have made efforts to analyse research output and funding at disease level, but only for a selection of diseases. Evans et al compared research output between countries for 19 disease groups, based on the International Classification of Disease (ICD)-9 chapters [5]. Gillum et al [6] and Gross et al [7] analysed burden of disease and research funding for a selection of 29 conditions, derived from the ICD. In various other studies funding, research output and burden of disease were described for specific diseases in a case by case approach. This was done for example for yellow fever [8] and neglected tropical diseases [9]. In other studies, total biomedical research output was analysed for specific countries [10, 11] or compared between countries [12, 13].

Text mining techniques are increasingly applied to biomedical text to uncover unseen relationships [14]. In this study we use these techniques to create a reference structure of disease groups and to catalogue publications accordingly. This opens a bridge between biomedical research output and other information available at disease level, which can contribute to a better understanding of the societal impact of biomedical science.

METHODS

Selection of biomedical publications

The analysis was based on the Clarivate Analytics WoS database available at the Centre for Science and Technology Studies (CWTS) of Leiden University. Since the goal of this study is to quantify research output by disease, we included biomedical research fields only. Of the 250 WoS research fields, we selected the 84 fields that are most medically oriented. We validated the selection by looking at the research output of the eight Dutch university medical centres: over 98% of their publications were in one of these fields. Appendix 1 provides a full list of research fields included in this study. The dataset was compiled in June 2014. It includes all publications in the 84 selected research fields, published between 2000 and early 2014, with WoS document type ‘article’ or ‘review’. Not all publications from the first six months of 2014 were available, due to periodical updating of the CWTS in-house version of the WoS database. The dataset contained 6.5 million publications in total.

Classification of publications by disease group

We defined 269 disease groups, based on the ICD-10 classification and covering the full spectrum of this classification. We used a two-step approach to categorise publications to disease groups.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

For peer review only CONFIDENTIAL

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

First, we categorised the author keywords listed by authors in their publications. In total, 158,700 unique author keywords were used in at least ten publications in our dataset. Of these keywords, the 32,400 most frequently used keywords (used in more than 70 publications each) were short listed and further evaluated. 21% of these keywords were specific for a single disease group. For example, the keyword 'Alzheimer's disease' was linked to 'dementia'. Many keywords were not suitable to use for categorisation to disease groups because they were either too general or not disease-specific. Examples of keywords not linked to a disease group are 'inflammation' and 'keyhole surgery'. We note that not all publications include author keywords.

In the second step, a text mining algorithm was used to search for disease-specific terms in titles and abstracts of publications. In this step, first a list of 10,983 unambiguous, disease-specific terms was generated by hand by medical professionals to characterize specific disease groups. Examples of terms for the disease group 'malignant neoplasm prostate' include 'prostate cancer', 'prostate carcinoma', 'malignant tumor prostate', and 'sarcoma prostate'. The generated disease-specific terms were then reviewed by another medical professional for ambiguity. Subsequently publications with one of these 10,983 terms in either title or abstract were assigned to the corresponding disease group. If the same publication was assigned to multiple disease groups, it was fully counted for all of them.

The method was validated in several ways. The first step was a manual examination of a random sample of 680 publications assigned to a disease group. Subsequently, a random sample of 315 publications not assigned to a disease group was manually examined. The examination was executed by research professionals among whom research coordinators and a clinical librarian of the Dutch university medical centres. The percentage of publications that could be assigned to a disease group was compared between WoS research fields. In addition, several institutional profiles resulting from the classification of research output to disease groups were discussed with researchers and deans from those institutions.

Classification of publications by institution and country

The name of an institution is often reported in many different ways in publications. Some authors for example report an abbreviated name while others report the full name, and some authors report the name of the university with which a hospital is associated while other authors report only the name of the hospital itself. These inconsistencies are problematic when analysing the research output of institutions. We addressed this problem by relying on the categorisation of affiliations used in the CWTS Leiden Ranking 2014 [15]. In this way we could compare the research output of the 750 largest universities worldwide (based on number of publications in WoS), of 1099 hospitals, and of 46 public research organisations. Publications from all affiliations, also those not included in the selected institutions, were included when comparing research output between countries.

Publications were assigned fractionally to institutions and countries. This was done based on the number of addresses in the address list of a publication in which a certain institution or country is mentioned. For instance, if a publication includes five addresses and two of these addresses mention Leiden University (e.g., two different departments within Leiden University), the publication is assigned to Leiden University with a weight of 2 / 5 = 0.4. So

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

the publication is not counted as a full publication for Leiden University but as 40% of a full publication. This methodology is known as address-level fractional counting [16].

Indicators of quantity and quality of research

We used several indicators of quantity and quality of biomedical research per disease group to provide quantitative insight in the research strengths of specific institutions and countries. Quantity was measured by the fractionally counted volume of publications of an institution or country. Citations are often seen as an indicator of scientific impact, or somewhat less precisely, as an indicator of quality. Since research fields differ in citation practices, comparison of citation counts between fields is difficult. Likewise, comparison of citation counts between older and more recent publications is problematic, because older publications have had more time to accumulate citations. To overcome this, we identified for each combination of a disease group and a publication year the 10% most cited publications globally. We used the volume of these 'top publications' as an indicator of quality of output when comparing countries or institutions. Only publications that appeared between 2000 and 2012 were used to identify 'top publications', since publications after 2012 were too recent for the calculation of meaningful citation statistics in 2014. Self-citations, that is, citations given by an author to his or her own work, were excluded. For the comparison of research portfolios between countries, between institutions, and over time, we used an institution's (or country's) share in the global publication volume per disease group as an indicator of the total volume (quantity). Additionally, we used the share of top publications in the total output of an institution (or country) as a size-independent indicator for quality. This relative measure enables a comparison of research output for different disease groups within the research portfolio of an institution (or country).

Patient and Public Involvement

No patients or public were involved in our study.

RESULTS

This section first describes the results of the validation of our method. Second, results for several applications of the method are described.

Validation of the proposed method

We were able to relate 54% of all publications in the selected 84 research fields to a disease group, 3.2 million publications in total. Of all publications, 29% were assigned to a single disease group, 14% to two disease groups, and 11% to three or more disease groups. Fields of research with a large share of disease-specific publications were mainly clinical research fields. Over 80% of all publications in research fields such as allergology, rheumatology, and clinical neurology were linked to a disease group. Research fields like ethics, microscopy, and biophysics had a much lower percentage of disease-specific publications (10%, 17%, and 27%, respectively). In these fields, we indeed would not expect a large share of the publications to be linked to a disease group, so the low percentages confirm that our method behaves as expected. We refer to appendix 1 for an overview of the share of disease-specific publications per research field.

Between 2000 and 2012, the annual volume of publications within the included research fields increased by 64%. In the same period, the volume of disease-specific publications increased by 92%. This means that disease-specific publications grew in share: from 48% in

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

CONFIDENTIAL For peer review only

2000 to 57% of total volume in 2012. After manual verification, we found that 2% of the sample of disease-specific publications (n=680) were incorrectly assigned to a disease group, and 1% of the sample of uncategorised publications (n=315) were incorrectly not assigned to a disease group, both indicating the method to be accurate. Incorrect links were mainly due to sentences such as “patients with diabetes were excluded” in the abstracts of publications.

About 1900 institutions were analysed in this study. Together these institutions accounted for 69% of the address lines in disease-specific publications worldwide. 78% of the disease-specific publications had at least one author from one of these institutions.

As expected, we found strong differences in the share of disease-specific publications between different types of research institutions in the Netherlands. We verified institution-specific results with researchers and deans of five top ranking institutions in the Netherlands and abroad. In all cases the disease-specific research output was in line with their expectations about their own institution’s position in relation to other institutions worldwide.

Application 1: Biomedical research output by disease group

Using our method, we can compare the research output between disease groups. The number of publications in the period 2000-2012 varies widely between disease groups, as shown in figure 1. ‘Other infectious diseases (not including HIV and tuberculosis)’ was the disease group with most publications. ‘Diabetes mellitus’, ‘metabolic diseases’, and ‘mood disorders’ were also large. The number of publications on malignant neoplasms was just a little bigger than the total publication volume on heart diseases.

[FIGURE 1]

Interestingly, the worldwide research profile by disease is not constant over time. Some disease groups have seen a rapid growth in research output, while other disease groups have grown only mildly in research output, as shown in figure 2. Lifestyle diseases (obesity and diabetes), cancers (lung, prostate, colon and breast), and mental disorders (depression and other mental disorders) gained in share in the worldwide research portfolio. On the other hand, diseases such as anaemia, pain in chest and throat, leukaemia, and HIV show a decreasing share in the total research portfolio, although the research output has still grown in absolute volume.

[FIGURE 2]

Application 2: Biomedical research output by disease by country

The most cited disease-specific research publications originate from a small set of countries. Figure 3 shows the relative share of countries in the 10% most cited publications per disease group. The top ten countries with the largest share in top 10% most cited research output account for 83% of the total body of disease-specific publications worldwide. Notably, the USA accounts for 45% of the top 10% most cited publications. There are however differences in research profiles between countries. For instance Canada has equal shares in top publications on ‘depression’ and ‘stroke’, while China has twice as many top publications on ‘stroke’ compared to ‘depression’.

[FIGURE 3]

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

It is possible to evaluate the development over time of each country's share in publication volume for a specific disease. Figure 4 shows the growth in number of breast cancer publications by country between 2000 and 2012. Although the number of publications of every country has grown during this period, some countries have grown faster than others. Most western countries have grown slower than the world average. Countries that have grown faster than average are mainly BRIC countries, with China showing 700% growth. Notably, Iran experienced a remarkable 3,500% growth in research output, but its total volume of disease-specific publications remains small.
[FIGURE 4]

Application 3: Research output by disease on an institution level

Our method allows for identification of institutions with a remarkable position in research on a specific disease group. We use Multiple Sclerosis (MS) as an example, but figure 5 can easily be constructed for all 269 disease groups used in this study. The figure shows for all institutions their volume of MS publications and their respective share in the top 10% most cited publications about MS worldwide. Harvard's unique position in MS research is illustrated by the facts that Harvard had the largest share in the total MS publication volume and that one in four of its publications was in the top 10% most cited publications about MS. Other centres with remarkable quantity and quality of MS research were University College London and VU University Amsterdam. A display like figure 5 recognises institutions that have a high quality without a high production.
[FIGURE 5]

Using our method it is possible to follow the research output of individual institutions for specific disease groups over time. As an example, figure 6 shows the rise of South African research output on HIV. Between 2000 and 2004, the annual South African research output on HIV is relatively constant, but from 2005 onward, several South African universities have grown rapidly, passing several famous HIV research institutions in volume. This growth seems partly due to growth of international collaboration. For instance, 10% of all South African publications on HIV were co-authored with Harvard University in 2012, while this was only 2% in 2005. During this time, internationally renowned Harvard scientists such as Bruce Walker and Till Barnighausen have started working part time for the University of Kwazulu-Natal.
[FIGURE 6]

In addition to comparing institutions for a specific disease, our method also allows us to map research portfolios of countries or institutions by disease, based on volume and top 10% publications. Using these portfolio maps, we can now compare complete disease-specific research portfolios between institutions. As an example, we plotted portfolio maps of four universities in figure 7. Substantial differences in their profiles can easily be seen. Harvard University has much larger publication volumes than the three others. Imperial College has a large number of disease groups with at least 30% of their publications counting as top publications. Both University of Amsterdam and Karolinska Institute have a remarkable position in research on malignant oesophageal neoplasms, whereas Imperial College does not.
[FIGURE 7]

DISCUSSION

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Our proposed method allows for systematic classification of publications in WoS to disease groups. We were able to classify 54% of all 6.5 million biomedical publications in the WoS database to a disease group. Between 2000 and 2012, 'Other infectious diseases' was the largest disease group with 337,485 publications. In this period, lifestyle diseases, cancers, and mental disorders have grown most in research output. On a country level, the USA was responsible for the largest number of top 10% most cited publications per disease group, with a global share of 45%. Iran (+3,500%) and China (+700%) have grown most in research volume. On an institution level, we were able to relate 78% of biomedical publications to a specific institution. Below we describe some examples of potential use and then discuss possibilities for future research.

Potential value of the proposed method

The method can be used for evaluation of historic research performance at the level of specific diseases. It can support decision making in management of research portfolios, showing relative strengths and weaknesses of institutions and countries. Combining these insights with indicators of innovation and research productivity [17] can illustrate whether research performance is aligned with successful transfer of scientific knowledge to clinical practice.

Linking the disease-specific research output to burden of disease provides insights in 'white spots' in global and regional research [18]. These insights can support fact-based allocation of research funding, making it possible to better align research portfolios to local or global needs and to adjust portfolios to changes of those needs over time. This can be the starting point for further understanding of what drives research output other than burden of disease, for instance; economic strengths, political structures, research legacy, etc. Quantitatively unravelling the different drivers that determine disease-specific publication volume could provide insights in how we can realign research efforts across countries to have greater impact on reduction of disease burden.

Opportunities for additional research

Using disease groups based on the ICD-10 classification has the advantage of being exhaustive: all diseases can be included. When looking for research on a rare disease, the used classification system is not specific enough. However, our method can be adapted to answer such specific questions by using specific author keywords and tailor-made text phrases to look for in titles and abstracts. Addition of MeSH descriptors next to author keywords can further complete the method, although this requires the use of other bibliographic databases, since WoS does not include MeSH descriptors. Ultimately, the use of dynamic and customised research categories will make it easier to find the institutions with the strongest positions in research on specific diseases, thus answering portfolio questions in ways that are not possible yet.

Our method classifies each publication to disease nomenclature but does not categorise the nature of disease-specific research. For example, a publication classified to a disease group could describe a new gene involved in the pathogenesis, analyse the societal impact of the disease, or merely state the disease as a potential application for a new surgical technique. Ideally, the method should be supplemented with additional categories that, based on text mining, can identify the type of research and application. Also clinical trial registers (e.g. https://www.clinicaltrialsregister.eu/ or https://clinicaltrials.gov/) can be included. As an

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

example, using a simple algorithm based on MeSH descriptors, it is possible to identify cell-based, animal-based, and patient-based research [19].

Now that publications are allocated to disease groups, bibliometric indicators of research quantity and quality can be combined with other information available on disease level. For instance, quality of care, patient reported health outcomes, cost of treatment, and patents. This can be valuable in aligning research and health care portfolios of university medical centres.

Conclusion

We have shown that it is possible to systematically link research output to disease groups. Our method makes it possible to compare research output by countries or institutions and to monitor changes in biomedical research output over time or by disease. The novelty and value of the method is that it allows a disease-specific analysis, for instance making it possible to compare research output with burden of disease. Since the major goal of biomedical research is alleviation of disease burden, our method allows for evaluating current strengths and shortcomings.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests disclosed

We have read and understood BMJ policy on declaration of interests and declare no competing interests.

Individual contributions

LL and NH made the definitions of disease groups, categorised the author keywords, and made the disease-specific keywords. TK, NH and LL performed the analysis. LL wrote the manuscript together with NH and TK. NH and LL validated the results with researchers and deans. LW implemented the text mining algorithm, assigned the publications to disease groups, and calculated the bibliometric statistics. The Centre for Science and Technology Studies (CWTS) at Leiden University provided the cleaned address data for the universities, hospitals and public research organisations included in the study. IM, LW and AG provided feedback on the manuscript.

Acknowledgements

The authors would like to thank the research coordinators and deans of the Dutch university medical centres for their contribution to the validation of this research method, the group of medical interns for their assistance in drafting the disease-specific terms, and prof. dr. Marcel Levi for his comments on the method.

Data sharing statement

Technical appendix can be provided. The appendix includes a definition of biomedical research by WoS research fields.

REFERENCE LIST

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

1. www.ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges/, date accessed: February 2016.
2. Røttingen JA, Regmi S, Eide M, et al. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *Lancet* 2013:382(10):1286-1307.
3. Thomson Reuters. Web of Science. http://www.webofknowledge.com, date accessed: February 2016.
4. Lipscomb C. Medical Subject Headings (MeSH). *Bull Med Libr Assoc.* 2000:88(3):265-266.
5. Evans J, Shim J, Ioannides J. Attention to local health burden and the global disparity of health research. *PLoS ONE* 2012:9(4):e90147.
6. Gillum L, Gouveia C, Dorsey E, et al. NIH Disease funding levels and burden of disease. *PLoS ONE* 2011;6(2):e16837.
7. Gross CP, Anderson GF, Powe NR. The relation between funding by the National Institutes of Health and the burden of disease. *N Engl J Med* 1999:340:1881-1887.
8. Bundschuh M, Groneberg D, Klingelhoefer D, et al. Yellow fever disease: density equalizing mapping and gender analysis of international research output. *Parasites and Vectors* 2013:6:331-43.
9. Adams et al. Thomson Reuters Global Research Report, 2012.
10. Minet Kinge J, Roxrud I, Volsset SE, et al. Are the Norwegian health research investments in line with the disease burden? *Health Res Policy Syst.* 2014:12:64.
11. Lascurain-Sánchez ML, García-Zorita C, Martín-Moreno C, et al. Impact of health science research on the Spanish health system, based on bibliometric and healthcare indicators. *Scientometrics* 2008:77:131
12. King D. The scientific impact of nations. Nature 2004:430:311-316.
13. Moses III H, Matheson D, Cairns-Smith S, et al. The anatomy of Medical Research, US and international comparisons. *JAMA* 2015:313(2);174-189.
14. Rodrigues-Esteban R. Biomedical text mining and its applications. *PLoS Comput Biol.* 2009 Dec;5(12):e1000597.
15. Waltman L, Calero-Medina C, Kosten J, et al. The Leiden ranking 2011/2012: Data collection, indicators and interpretation. *J. Am. Soc. Inf. Sci. Technol.* 2012:63(12):2419-2432.
16. Waltman L, Van Eck N. Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of informetrics* 2015:9(4):872-894.
17. Balas EA and Elkin PL. Technology Transfer from biomedical research to clinical practice: measuring innovation performance. *Eval Health Prof.* 2013;36(4):505-17.
18. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015. http://dx.doi.org/10.1016/S0140-6736(15)60692-4.
19. Weber G. Identifying translational science within the triangle of biomedicine. *J Transl Med* 2013;11:126-36.

BMJ Open: first published as 10.1136/bmjopen-2017-020818 on 4 June 2018. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

For peer review only

CONFIDENTIAL