City Research Online

City, University of London Institutional Repository

This is the published version of the paper.

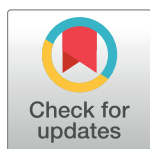This version of the publication may differ from the final published version.

The proportion of cancer-related entries in PubMed has increased considerably; is cancer truly "The Emperor of All Maladies"?

Constantino Carlos Reyes-Aldasoro*

School of Mathematics, Computer Science and Engineering, City, University of London, London, United Kingdom

* reyes@city.ac.uk

Abstract

In this work, the public database of biomedical literature *PubMed* was mined using queries with combinations of keywords and year restrictions. It was found that the proportion of Cancer-related entries per year in PubMed has risen from around 6% in 1950 to more than 16% in 2016. This increase is not shared by other conditions such as AIDS, Malaria, Tuberculosis, Diabetes, Cardiovascular, Stroke and Infection some of which have, on the contrary, decreased as a proportion of the total entries per year. Organ-related queries were performed to analyse the variation of some specific cancers. A series of queries related to incidence, funding, and relationship with DNA, Computing and Mathematics, were performed to test correlation between the keywords, with the hope of elucidating the cause behind the rise of Cancer in PubMed. Interestingly, the proportion of Cancer-related entries that contain "DNA", "Computational" or "Mathematical" have increased, which suggests that the impact of these scientific advances on Cancer has been stronger than in other conditions. It is important to highlight that the results obtained with the data mining approach here presented are limited to the presence or absence of the keywords on a single, yet extensive, database. Therefore, results should be observed with caution. All the data used for this work is publicly available through PubMed and the UK's Office for National Statistics. All queries and figures were generated with the software platform Matlab and the files are available as supplementary material.

Introduction

The database *MEDLINE* of the United States National Library of Medicine (NLM) and its search engine *PubMed* (https://www.ncbi.nlm.nih.gov/pubmed) have grown to include over 26 million entries: 26,710,394 on the 30 November 2016. In MEDLINE, each entry is indexed with Medical Subject Headings (MeSH) and various field descriptors such as author, date, title, publication type, etc. For a complete list of the MEDLINE elements, the reader is referred to https://www.nlm.nih.gov/bsd/mms/medlineelements.html. These fields allow specific searches to be performed in PubMed by restricting the search to one, several or all fields and

logical combinations with operators such as AND, OR, NOT are available as well. Whilst PubMed it is not without criticism for its inconsistency of terminology [1,2], document ranking not content-based [3], and "unwelcoming complexity" of its searches and terms [4], PubMed is considered to be "the most widely used database of biomedical literature" [3] and it was found to have better precision than Google Scholar (http://scholar.google.com) [5]. An interesting discussion of advantages and disadvantages of PubMed is summarised in a website titled "*Twenty million papers in PubMed*: *a triumph or a tragedy*?" [6].

This work was motivated by the interest on the presence of Cancer-related publications in PubMed. Within the database of PubMed, more than 3 million entries correspond to Cancer, which correspond roughly to 12% of the total entries. However, the proportion has increased significantly from around 6% in the 1950 to 16% in 2016. To investigate this increase and its possible causes, this work generated data mining tools to perform a systematic mining of Cancer-related terms in PubMed. The platform selected for the mining was Matlab® (The Mathworks ™, Natick, USA) which is a widely-used programming environment in Science and Engineering. All the code used for this publication is provided in S1 File. Matlab can be used to retrieve the information and further analyse or display the data with its powerful tools, without the need of specialised web-based tools like MeSHy [3] or GoPubMed [7]. Furthermore, data from other websites like the UK's Office for National Statistics (https://www.ons.gov.uk) was also extracted and displayed with Matlab for this work. The results are essentially limited as they are restricted to one database and the presence or absence of the keywords used in this study.

Materials and methods

MEDLINE PubMed was queried with a combination of keywords using the software platform Matlab® (The Mathworks ™, Natick, USA) and displayed directly with Matlab. The keywords were combined to form a Uniform Resource Locator (URL), which started with the basic address of PubMed 'https://www.ncbi.nlm.nih.gov/pubmed/?term=', to which the keywords to be searched in '[All Fields]' or '[MeSH Terms]' were added. URLs do not accept special characters like spaces, brackets [] or quotes ' " and these need to be converted to the ASCII character set. For instance, the following PubMed search "Cancer"[All Fields] has to be translated to the following URL: https://www.ncbi.nlm.nih.gov/pubmed/?term=%22Cancer%22%5BAll+Fields%5D.

A series of hypothesis were investigated by adding keywords to the URL. Queries on the terminology related to Cancer were investigated with the keywords: *neoplasms*, *cancer*, *tumor*, *neoplasm*, *tumors*, *oncology*, *metastasis*, *cancers*, *tumour*, *tumours* and *neoplasia*. *The growth of Cancer was compared against other conditions*, *namely*: *AIDS*, *Malaria*, *Tuberculosis*, *Diabetes*, *Cardiovascular*, *Stroke* and *Infection*. *Queries for organ-related cancers were investigated with the keywords*: *Bladder*, *Bowel*, *Brain*, *Breast*, *Kidney*, *Leukaemia*, *Liver*, *Lung*, *Lymphoma*, *Melanoma*, *Mouth*, *Ovarian*, *Pancreas*, *Prostate*, *Sarcoma*, *Stomach*, *Testicular*, and *Uterus*. Year-on-year variation on Cancer and other conditions were queried by restricting the date of publication to a single year.

The URL was passed as an argument to the Matlab command '*urlread*', which read and returned a variable with the webpage that PubMed produced for that specific query. The webpage was stored in a string variable as a sequence of alphanumeric characters. The variable was then searched for the string 'count' which provided the number of entries contained in that search. To restrict the search to one particular year, e.g. 1990, the following term was added to the search: 'AND "1990"[DP]' where 'DP' is the MEDLINE Field for 'Date of Publication'. These yearly queries were used to perform searches between 1950 and 2016.

The code to generate the queries and the figures is included in S1 File. There were no restrictions of language, country or publication type. However, if necessary, it would be easy to use these MEDLINE field with the source code provided with this work.

Results and discussion

Terminology

To investigate the variability of terminology [1] and in particular of Cancer keywords [2], several keywords and some of their combinations were explored. The search indicated that the Cancer-related keyword with more entries was '*neoplasms*' with more than 2 million entries followed by *'cancer', 'tumor', 'neoplasm', 'tumors', 'oncology', 'metastasis', 'cancers', 'tumour', 'tumours'* and *'neoplasia'* (Fig 1). It was interesting to notice that whilst there are more entries for '*neoplasms*' in plural than for '*neoplasm*', the opposite was true for '*tumor/tumors*', '*cancer/cancers*' and '*tumour/tumours*' where the term in singular resulted in a larger number of entries. It would have been expected that a term in singular would include the plural, but for neoplasms this was not the case.

The overlap of terms for a single entry was explored by performing pair-wise searches between all keywords: i.e. '*neoplasms* OR *cancer*', '*neoplasms* OR *tumor*', etc. Partial results are shown in Fig 2a as bars indicating the number of combined entries. Since these searches are symmetric, only one case is shown with the diagonal indicating the results with the single entry (i.e. *tumor* OR *tumor*). Whilst it was expected that combinations of distinct keywords would report an increase of entries with the combination (e.g. *neoplasms* = 2,271,532, *cancer* = 1,742,236, *neoplasms* OR *cancer* = 2,909,324) it was curious to notice that the combinations of singular and plural terms also increased the number of entries (e.g. *tumor* = 1,349,097, *tumors* = 510,353, *tumor* OR *tumors* = 1,543,897. These results indicated that there were entries, which used one, the other or both keywords.

The repetition of the keywords was explored by changing the term OR for AND, i.e. '*neoplasms* AND *cancer*', '*neoplasms* AND *tumor*'. Partial results shown in Fig 2b show a
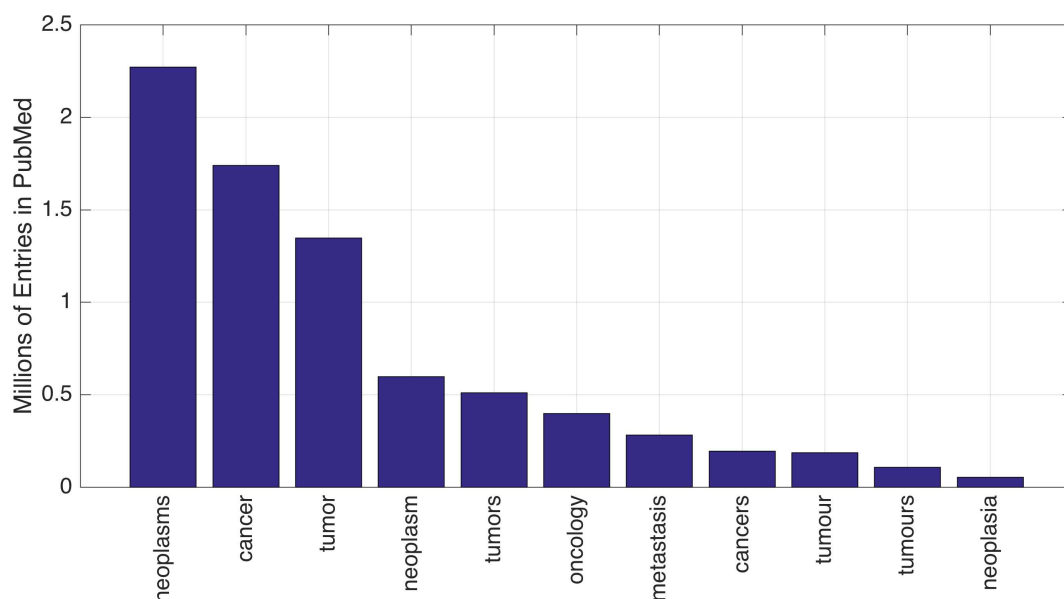
Fig 1. Number of cancer-related entries for different keywords listed in decreasing order.

doi:10.1371/journal.pone.0173671.g001
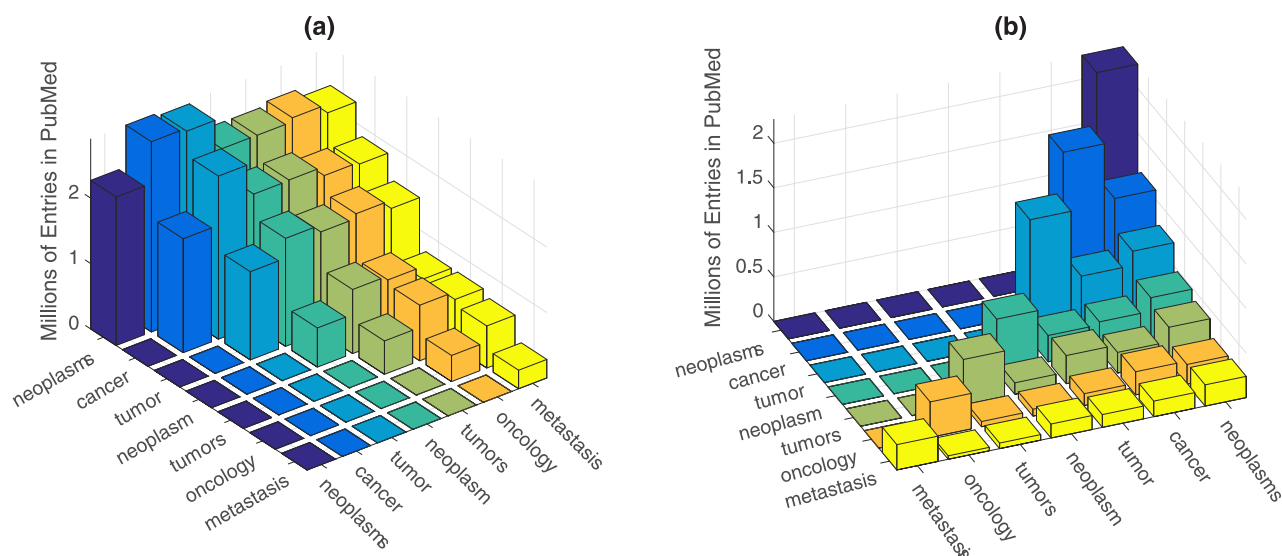
Fig 2. Number of entries in PubMed for searches with pairs of keywords. For all cases, each column represents the result for the combination of the pair of keyword on the two axes. (a) Combinations with the operator OR. (a) Combinations with the operator AND. The diagonal corresponds to a single keyword and since the matrix is symmetric a single side is shown.

considerable number of publications that contain both keywords. Again, the diagonal contains the result of a single keyword and only one combination is shown, as the search is symmetric. These results together those of Fig 2a indicate that single words are not sufficient to search for Cancer in PubMed. Another observation of these searches is that the number of entries with American spelling for *tumor* (1,347,989) was around sevenfold larger that its counterpart in British spelling *tumour* (187,719).

Year to year PubMed entries of cancer and other conditions

Next, year-to-year queries were performed for to investigate the number of PubMed entries that were indexed in a given year. The Cancer-related entries were as defined by a subset of the keywords previously mentioned (*neoplasms, cancer, tumor, tumour, oncology*) in all of the MEDLINE fields. These were compared against the total number of entries of a single year; in these cases, the query was restricted only to the date of publication. The proportion of Cancer-related entries to the total is shown in Fig 3a, where it can be seen a rise from 6% in 1950 to 16% in 2016. To compare this rise, the following conditions were also queried on a yearly basis: *AIDS, Malaria, Tuberculosis, Diabetes, Cardiovascular, Stroke* and the more general *Infection*. In the case of *AIDS*, the query was performed with the combination of the following keywords: "*acquired immunodeficiency syndrome*" [MeSH Terms] OR ("*acquired*" [All Fields] AND "*immunodeficiency*" [All Fields] AND "*syndrome*" [All Fields]) OR "*acquired immunodeficiency syndrome*" [All Fields] OR "*aids*" [All Fields]. The other conditions were sufficiently specific to use only one keyword. A zoom into the lower portion of the graph is shown in Fig 3b.

The trends shown in Fig 3 are quite interesting. First, none of the other conditions is close to Cancer in the number of entries in PubMed. Even the more general term Infection is more than 10% below in 2016. Second, the trend of growth of Cancer is way higher than any of the other conditions growing from 8% in the 1970 to 16% in 2016.

The proportion of AIDS-related entries rapidly increased from nearly zero to close to 2% in a few years in the 1980s. The peak in the late 1980s is close to the development of anti-retroviral treatments like *zidovudine* [8–10] and their approval, as a therapy against HIV, by the US

(a)

(b)
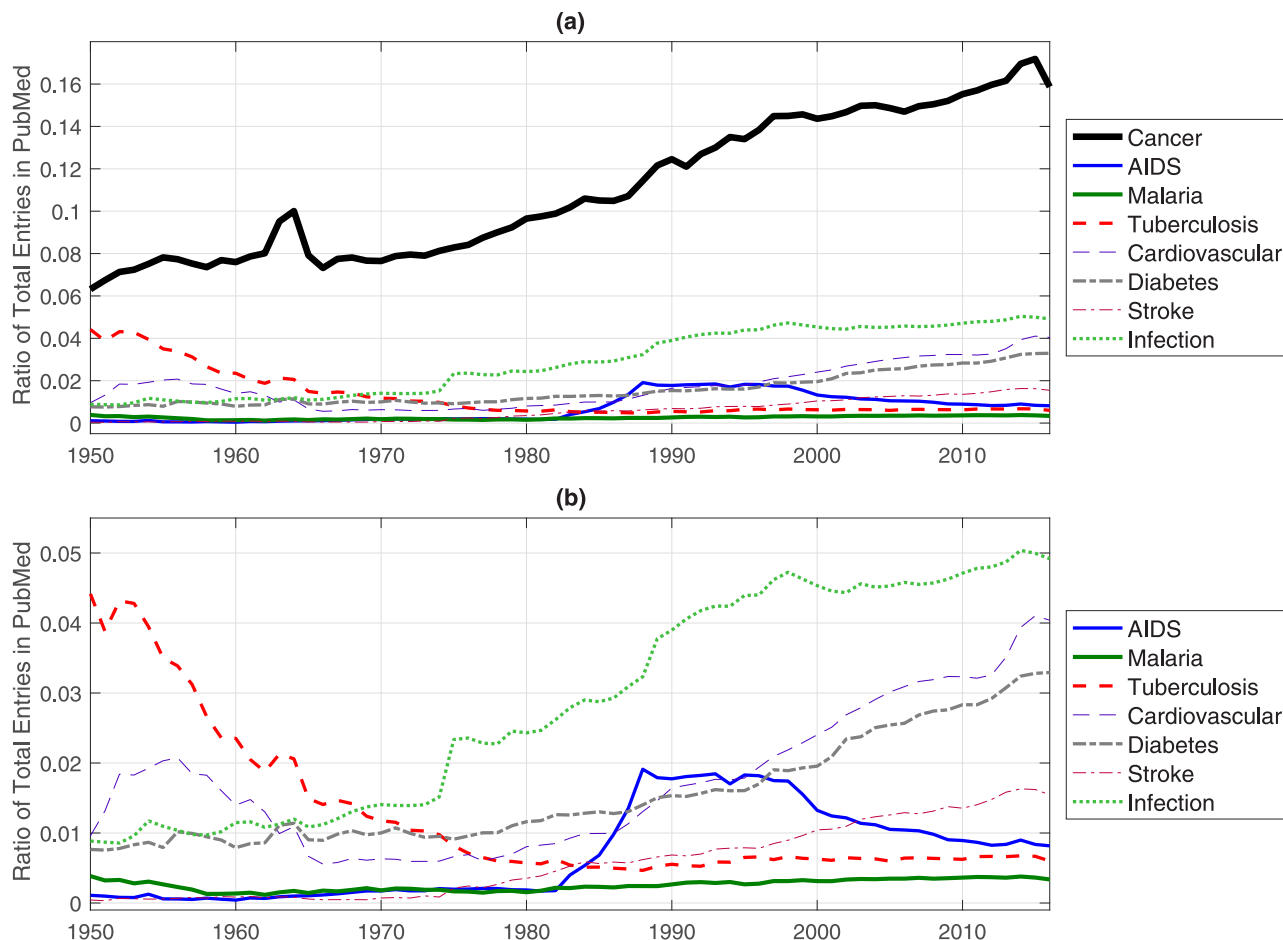
Fig 3. (a). Ratio of a series of condition-related entries in PubMed to the total number of entries per year. Notice how Cancer entries have increased from around 6% in the 1950s to 16% in 2016. All other conditions are considerably below Cancer. (b) Zoom into the lower values of the vertical axis of (a). Notice the different trends of each condition.

doi:10.1371/journal.pone.0173671.g003

Food and Drug Administration (FDA) (10,11) in 1987. The entries remained relatively constant until the late 1990s and after that date there has been a steady decline in the number of entries. According to a recent study, during this time the "effective number of infections" plateaued in the 1990s in the United States [11]. However, the deaths of HIV/AIDS were continued to rise to become one of the leading causes of death globally. According to Lozano [12], the fatalities caused by AIDS were 0.30 million in 1990, 1.5 million in 2010, and reached a peak of 1.7 million in 2006. Clearly, the number of entries indexed by PubMed, has diverted to other areas.

Even when *Cardiovascular* disease is considered "the world's top killer" [13,14], the contribution in PubMed oscillates between 0.5% - 4%. Interestingly, it was close to 2% in the 1950s, then dropped and remained at a low level until the 1980s, from which it has grown steadily with a very recent sharp increase. The decreasing trend may be related to a decrease in mortality rate, probably due to the decrease on some risk factors like smoking, and high blood pressure. In the United States the rates of death from coronary heart disease peaked in 1968 [15]. The increase experienced since the 1990s may be related to the sharp increase in two other risk factors: obesity and diabetes [16].

The entries related to the term *Infection*, have risen from 1% in the 1950s to around 5% in 2016. Whilst this is a significant increase and a position higher than most other conditions queried in this work, it still does not correspond to ranking in worldwide mortality. According to World Health Organization [17,18] lower respiratory infections constitute the third leading cause of death and diarrhoeal diseases, most of which are infections of the intestines, are the sixth cause of death.

Diabetes and *stroke* have shown similar trends, from steady levels of 1% and near to 0%, respectively, between 1950 and 1980s, and then a steady increase to reach around 3% and 1.5% respectively, in 2016. The increase of research in both conditions may be related with what some reports consider "alarming increase" in obesity [15,19,20], as obesity together with sedentary lifestyles are associated with significant higher rates of diabetes [16,21,22] and obesity is a risk factor for stroke and one of its most common co-morbidities [23–26]. There were reports of an "obesity paradox" [27] where "in cardiovascular disease populations, obese patients survive better", however, this has been proven to be result of a biased selection [28]. As with other conditions, the proportion of research entries does not reflect the rank of global causes of death [12] of these conditions.

An interesting case is related to *Tuberculosis* (TB). It has been reported that the absolute number of research entries related to Tuberculosis has increased [29], however, as a proportion to the total research reported in PubMed, it has decreased from around 4% in the 1950, when it was the second after Cancer (Fig 3), to a lowest point of 0.5% in the late 1980s and then experiencing a minor increase, probably related to the United States (US) Government National Tuberculosis Elimination plan released in 1989 [30]. Even when the number of cases of TB reported in the US is relatively small, 9,951 in 2012, as compared with the global 8.6 million cases and 1.3 million deaths [31], there is significant concern as the majority of the TB reported cases in the US correspond to individuals born outside the US, thus reducing TB overseas would reduce the domestic cases [32].

Despite its significant contribution of global causes of death [12], the ratio of entries related to *Malaria* is below 0.5% for the whole period of analysis. The reasons behind this low level must be numerous and complicated. One report observed that research investment in Malaria followed colonial ties between the United Kingdom and former colonies [33]. Sometimes, it is considered that investment in drug development, especially from the private sector, is dedicated exclusively to drugs that will be marketable and profitable in the developed world [34,35]. This has led to some diseases, like Malaria to be considered as "neglected" [36] and others like sleeping sickness and Chagas as "most neglected" [34]. There has been progress with funding from public-private initiatives, notably *The Global Fund to Fight AIDS, Tuberculosis, and Malaria* [37]. However, it has been observed that the funding for the control of Malaria is still inadequate [38] and that efficient prioritisation of resources could save greater number of lives [39].

Finally, the combination of all the previous terms accounted to 31.42% of the entries indexed on 2016. To investigate the remaining 68.58% a series of queries with keywords from the MESH terms were performed. The results are presented as percentages in brackets after the keyword: *Adrenal (0.23%), Aggression (0.15%), Alcohol (1.07%), Allergy (0.79%), Alzheimer's (0.68%), Antigen (0.77%), Behavior (3.74%), Chemistry (8.69%), Dementia (0.61%), Enzyme (2.13%), Epidemiology (4.74%), Epilepsy (0.48%), Fracture (0.75%), Fungal (0.63%), Geriatric (0.45%), Glaucoma (0.23%), Hepatitis (0.66%), Hormone (0.87%), Inheritance (0.13%), Injections (0.51%), Insurance (0.45%), Labor (0.28%), Lactose (0.06%), Leprosy (0.05%), Life Expectancy (0.11%), Lupus (0.25%), Macular Degeneration (0.12%), Magnetic Resonance (1.95%), Microscopy (2.05%), Mitochondrial (1.03%), Morgue (0%), Neurology (2.63%), Nutrition (2.11%), Nursing (2.04%), Osteoporosis (0.29%), Parasitic (0.27%), PCR (1.74%), Plaque (0.34%),*

Poisoning (0.18%), *Polymers (0.52%)*, *Phobia (0.03%)*, *Therapy (8.39%)*, *Tomography (1.92%)*, *Toxins (0.22%)*, *Vasectomy (0%)*, *Vitamin (0.66%)*, *Wound healing (0.33%)*. It should be noticed that only 2 of the extra 47 keywords were above 5%: "Chemistry" with 8.69% and "Therapy" at 8.39%. Another important observation is that there may be entries that are related to more than one keyword and would appear in more than one query. The total for these 47 keywords is 56.32%.

Year-to-year PubMed entries of organ-specific cancer queries

The year-to-year results can be further explored by using organ-specific keywords to investigate the trends related to some specific Cancers. The following 18 keywords were used to explore some common Cancers: *Bladder, Bowel, Brain, Breast, Kidney, Leukaemia, Liver, Lung, Lymphoma, Melanoma, Mouth, Ovarian, Pancreas, Prostate, Sarcoma, Stomach, Testicular, Uterus*. Furthermore, the results were ranked by the difference between the percentages in the periods 1950–1955 and 2011–2016 to investigate if the use of a specific keyword had increased or decreased. The results with the highest increase are shown in Fig 4a, the highest decrease in Fig 4c and intermediate results in Fig 4b. The highest increase was related to the keyword *Breast* from 4.5% to nearly 11% of the cases, followed by *Liver* and *Prostate*. The highest decrease was for the keyword *Uterus* from 6% to 0.2% followed by *Sarcoma* and *Stomach*. It is important to highlight that an increase or decrease in the number of entries does not necessarily imply an increase or decrease on the incidence of the organ-related Cancer; it may be that there was a change on terminology. This may be the case for *Uterus*, which experienced a dramatic drop between 1962 and 1963. Further investigation into that drop is beyond the scope of this publication.

Why have cancer entries risen in this way in PubMed?

As an attempt to unravel the reasons behind the increase, which could be considered as dominance, of Cancer-related entries in PubMed, several hypotheses were examined.

The first hypothesis to be tested was the following: *Cancer incidence is increasing and thus the research related to Cancer has increased*. The literature has a considerable number of reports in which the incidence of certain Cancers has increased, and others for which the incidence has decreased. In some cases, the *decrease* in incidence has been related to a geographic population: cervical cancer in Spain [40], renal cell carcinoma in Sweden [41], colorectal cancer in the United States [42], testicular cancer in Denmark [43]. In other cases, the reduction has been linked with changes of treatments, notably hormone replacement therapy [44,45] or the use of screening or testing like Prostate Specific Antigen (PSA) for prostate cancer in Italy [46]. On the other hand, *increase* of cancer of the tongue has been recorded in Nordic countries [47], thyroid cancer in the U.S. [48] and Australia [49], distant stage breast cancer in the US [50] and India [51]. Variations in screening and lifestyle factors are considered to provide substantial incidence variation [52] but others suggest variations of other risk factors [53].

As an examination of this hypothesis, Cancer in the United Kingdom (UK), as a subset of the global population, was investigated in a similar way as the mining of the previous figures. The incidence and mortality data was mined, again with Matlab from the Office for National Statistics (http://www.ons.gov.uk/). Whilst the incidence of all reported Cancers in the UK has slightly increased between 1995 and 2014, the growth is nothing comparable to the rise shown in Fig 3. Furthermore, the trend in mortality has decreased in the same period.

The second hypothesis is related to the funding: *more funding is available for Cancer research and therefore the proportion of publications has risen*. MEDLINE has a field for Grant Number [GR], which states the grant and the funder, for example: "*GR—R21 CA194661/CA/*

(a)

(b)

(c)
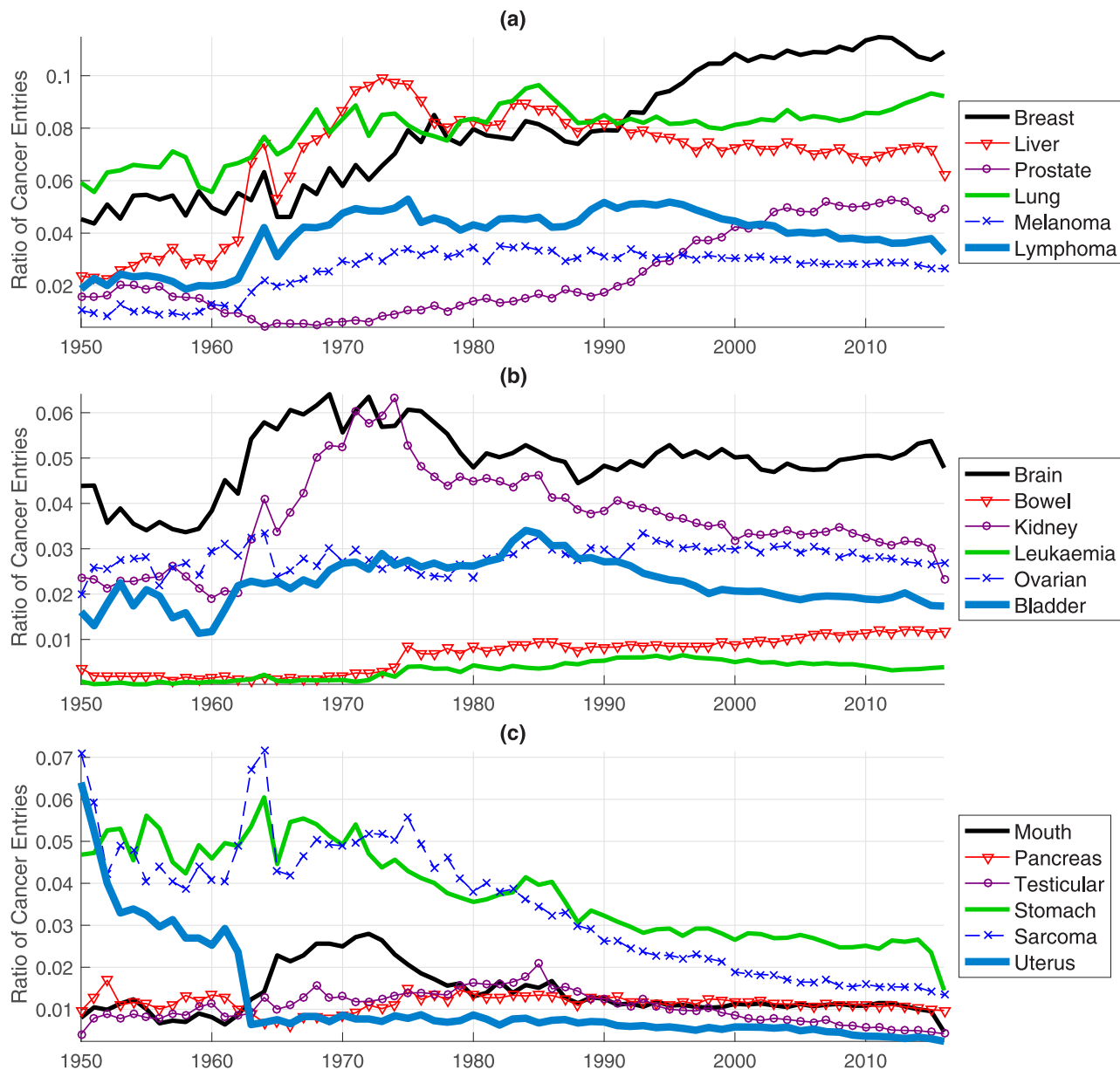
Fig 4. Ratios of the cancer entries related to organ-specific keywords. The trends have been ranked and presented according to (a) largest increase, (b) intermediate increase and (c) largest decrease from 1950s to 2016.

doi:10.1371/journal.pone.0173671.g004

NCI NIH HHS/United States". The National Cancer Institute (NCI) (https://www.cancer.gov/) is part of the National Institute of Health (NIH) (https://www.nih.gov/), which in turn is which is one of 11 agencies that compose the Department of Health and Human Services (HHS) (https://www.hhs.gov/) of the USA. The NIH is a major funder of research with more than 2 million entries that list NIH in the grant number field. Thus, PubMed was queried for the entries with grant numbers that contained NIH or NCI, this last one is a subset of NIH. The number of grants is rather small from 1950 until 1980 when a rise of numbers is present. From this date, the ratio of NCI grants to the NIH is relatively constant at around 20%
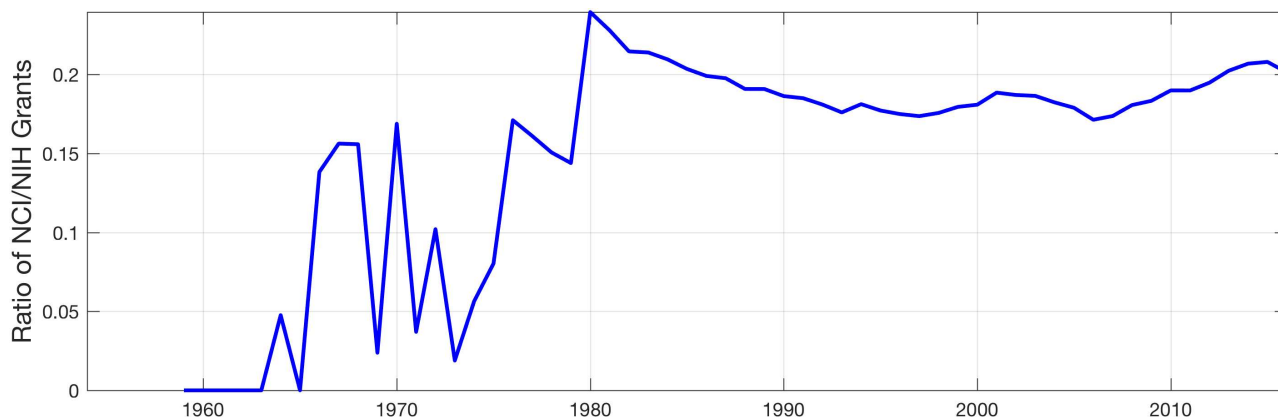
Fig 5. Ratio of the number of entries that report a grant number of the National Cancer Institute (NCI) over the number of entries that report a grant number of the National Institute of Health (NIH) of which the NCI is part. This ratio is an indication of the Cancer-funding from this Institute in the United States. It can be seen that the the ratio has been relatively constant at around 20% from 1980.

doi:10.1371/journal.pone.0173671.g005

suggesting, that as a proportion of its funded research, the NIH has not changed from the 1980s (Fig 5).

The last hypothesis to be tested was the following: *scientific advances have had a higher impact in Cancer than in other diseases*. To test this hypothesis, it was necessary to select the relevant scientific advances with a relevant keyword through which to mine them in PubMed. The first scientific advance selected was the advent of genetic-related research. PubMed was thus queried for all publications that contained the keyword *DNA*. The output was then separated into two groups, those entries related to *Cancer* as defined by all the keywords previously described, and those that did not and presented as ratios of the total (solid line in Fig 6). The figure shows that the proportion of Cancer-related entries rises from around 5% in the early 1950s to nearly 30% in 2016, with a noticeable jump in the mid 1980s.

The increase in this test seems to confirm the hypotheses that genomic advances have had a higher impact in Cancer than in other diseases, and consequently, at least one of the reasons behind the general increase of Cancer-related entries in PubMed. Whilst for Cancer genomic-related advances include the discovery of oncogenes (i.e. *Src* [54–56]), tumour suppressor
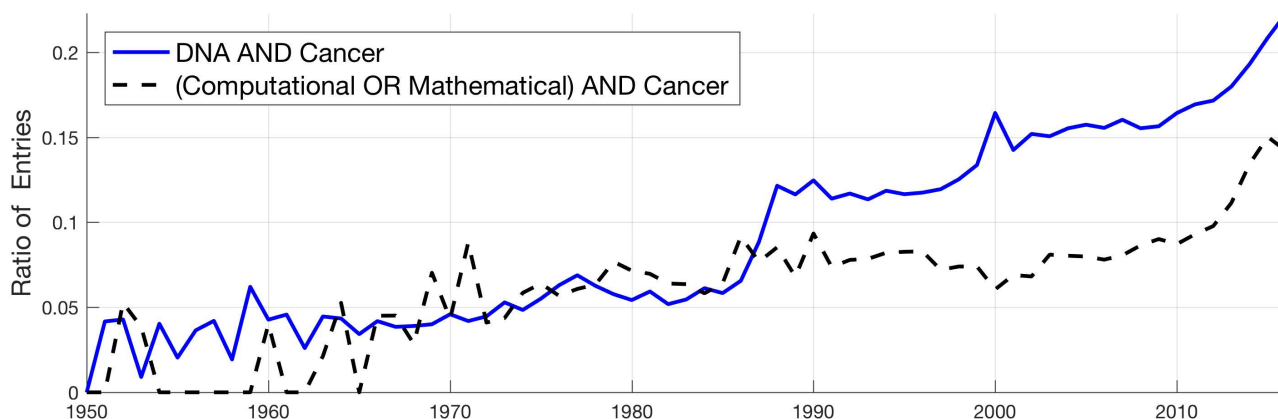
Fig 6. Ratios of all entries with the terms DNA and (Computational OR Mathematical), with and without cancer-related keywords as an indication of the impact that advances in these areas have had in cancer research. The ratio of Cancer has increased since the 1950s, with a particular surge in the mid 1980s for DNA.

doi:10.1371/journal.pone.0173671.g006

genes (i.e. *TP53* [57,58]) and mutations that confer well-known higher risks of cancer, e.g. *BRCA1*, *BRCA2* [59,60], the advances have not materialised so clearly for other conditions, like cardiovascular disease. It has been reported that few genome-wide studies have revealed "variants that, on their own, boost the risk of cardiovascular disease" [13]. Other trials showed that genetic tests did not show improved patient outcomes when observing patient's genetic variants and dosing regimens of the anticoagulant drug warfarin [61]. As such, Cancer-related predictive genetic tests can be now performed, for instance, by the National Health Service in the UK [62].

A second scientific advance, which was tested was on scientific areas that may seem initially unrelated to Life Sciences, namely, Computing and Mathematics. To test these, PubMed was queried for the keywords (*Computational* OR *Mathematical*). The output was separated again in two groups, those that were Cancer related and those that did not include the previously defined keywords and the ratio was calculated (dashed line in Fig 6). A similar trend as that of DNA was found for the terms computational and mathematical; the proportion of Cancer-related entries has grown, before the 1970 it was zero in some years, then around 6–8% in the period 1970–1990, and has had a significant increase to around 14%.

The impact of mathematics and computing on Life Sciences has been identified in several areas. Recently, the newly appointed director of the US National Institute of Mental Health (NIMH) declared "To appreciate and exploit that complexity (of the brain) we need to integrate everything we know, from molecular biology to behaviour, into the models of how the brain works. That requires serious math." [63]. There have been numerous publications in which different Cancer-related aspects have been investigated with mathematical and computational modelling, for instance: tumour vascular organisation [64], early-stage cancer interactions [65], tumour-induced angiogenesis [66,67], drug-transport in tumours [68,69], progression and aetiology of leukaemia [70,71], patient-specific drug treatments [72,73]. Some authors have stated that research related to Cancer has moved from solely experimental work to a combination with mathematical and computational modelling [74,75].

Conclusion

This work developed a series of data mining tools as files of the software platform Matlab, which are available in S1 File. The tools were used to investigate the presence of Cancer-related entries in PubMed. It was found that Cancer-related entries have grown substantially from 6% in the 1950 to 16% in 2016. The hypothesis that that scientific advances in the areas of Genetics, Computing and Mathematics had a stronger influence in Cancer than other areas was supported by the results, which indicated an increase in Cancer-related PubMed entries as a ratio of the total. A particular change in the trends was noticed in the mid 1980s, and could be related to a series of developments in genetics in that time: Genome sequencing [76], development of Polymerase Chain Reaction (PCR) [77] or gene mapping of a human disease [78]. However, a finer mining is necessary to correlate individual advances like these ones with its impact on Cancer-related research.

Other factors could also influence the results. Cancer awareness is rising through efforts to improve literacy [79] promotion in schools [80] or training of health workers [81]. Patients diagnosed with Cancer are likely to survive a considerable period of time through which they will increase awareness. Diseases like atherosclerosis, on the other hand, develop 'silently' over decades before any symptoms occur [82].

This work has several limitations, of which the most important one is that only one database, albeit a very large one, was mined. There may be a substantial amount of non-reported research in commercial companies related to Cancer and other conditions that do not appear

in PubMed. Second, the results depended on the presence or absence of keywords, some of these were examined in the section Terminology, but others were beyond the scope of this work, for instance, uterus/uterine. Third, there may be overlap of the results when one entry contains more than one keyword and thus the entries can be counted more than once.

Many other factors, which were not tested, could have an influence in the results presented. Besides funding from the NIH other sources of funding such as charities like *American Cancer Society*, *Cancer Research UK*, *National Foundation for Cancer Research* or *Worldwide Cancer Research* can have a significant impact in the research produced. The funding of pharmaceutical companies interested in developing cancer treatments may also be significant and have an impact on the research output, especially when compared against Malaria or other neglected diseases.

Cancer-related research has grown considerably. PubMed contains millions of publications, which describe techniques for diagnosis, management and treatment of Cancer. The oncologist Siddhartha Mukherjee wrote a comprehensive and captivating book titled "The Emperor of All Maladies, A Biography of Cancer" [83], which, interestingly, is not indexed in PubMed. At the beginning, within the Author's Note, Mukherjee states that his ultimate aim is to raise the questions "*Is cancer's end conceivable in the future*? *Is it possible to eradicate this disease from our bodies and societies forever*?" At this moment, the "Emperor" is the most prevalent disease in Europe [84], causes millions of deaths worldwide and is dominating the entries in PubMed. Whether, through the combined work of all authors behind those millions of publications, the Emperor can be ultimately contained and eradicated remains to be seen.

Supporting information

S1 File. DataMiningCode_PLOSONE_CancerInPubMed.zip. This compressed file contains the Matlab code that was used to query PubMed and to generate the figures of the paper. (ZIP)

Author Contributions

Conceptualization: CCRA.

Data curation: CCRA.

Formal analysis: CCRA.

Investigation: CCRA.

Methodology: CCRA.

Software: CCRA.

Visualization: CCRA.

Writing – original draft: CCRA.

References

1. Søgaard M, Andersen JP, Schønheyder HC. Searching PubMed for studies on bacteremia, bloodstream infection, septicemia, or whatever the best term is: a note of caution. Am J Infect Control. 2012; 40: 237–240. doi: 10.1016/j.ajic.2011.03.011 PMID: 21775021

2. Vanteru BC, Shaik JS, Yeasin M. Semantically linking and browsing PubMed abstracts with gene ontology. BMC Genomics. 2008; 9 Suppl 1: S10.

3.   Theodosiou T, Vizirianakis IS, Angelis L, Tsaftaris A, Darzentas N. MeSHy: Mining unanticipated PubMed information using frequencies of occurrences and concurrences of MeSH terms. J Biomed Inform. 2011; 44: 919–926. doi: 10.1016/j.jbi.2011.05.009 PMID: 21684350

4.   Abbasi K. Simplicity and complexity in health care: what medicine can learn from Google and iPod. J R Soc Med. 2005; 98: 389. doi: 10.1258/jrsm.98.9.389 PMID: 16140847

5.   Anders ME, Evans DP. Comparison of PubMed and Google Scholar literature searches. Respir Care. 2010; 55: 578–583. PMID: 20420728

6.   Hull D. Twenty Million Papers in PubMed: A Triumph or a Tragedy? «O'Really? http://duncan.hull. name/2010/07/27/pubmed-20-million/

7.   Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 2005; 33: W783–786. doi: 10.1093/nar/gki470 PMID: 15980585

8.   Riesenberg DE, Marwick C. Anti-AIDS agents show varying early results in vitro and in vivo. JAMA. 1985; 254: 2521, 2527, 2529. PMID: 2997490

9.   Ruzicka T, Fröschl M, Hohenleutner U, Holzmann H, Braun-Falco O. Treatment of HIV-induced reti- noid-resistant psoriasis with zidovudine. Lancet Lond Engl. 1987; 2: 1469–1470.

10. Richman DD. The treatment of HIV infection. AIDS Lond Engl. 1988; 2 Suppl 1: S137–142.

11. Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, et al. 1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature. 2016; 539: 98–101. doi: 10. 1038/nature19827 PMID: 27783600

12. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Bur- den of Disease Study 2010. Lancet Lond Engl. 2012; 380: 2095–2128.

13. Hayden C Erika. Cardiovascular disease gets personal. Nat News. 2009; 460: 940–941.

14. Gerszten RE, Wang TJ. The search for new cardiovascular biomarkers. Nature. 2008; 451: 949–952. doi: 10.1038/nature06802 PMID: 18288185

15. Ford ES, Ajani UA, Croft JB, Critchley JA, Labarthe DR, Kottke TE, et al. Explaining the Decrease in U. S. Deaths from Coronary Disease, 1980–2000. N Engl J Med. 2007; 356: 2388–2398. doi: 10.1056/ NEJMsa053935 PMID: 17554120

16. Hedley AA, Ogden CL, Johnson CL, Carroll MD, Curtin LR, Flegal KM. Prevalence of overweight and obesity among US children, adolescents, and adults, 1999–2002. JAMA. 2004; 291: 2847–2850. doi: 10.1001/jama.291.23.2847 PMID: 15199035

17. World Health Organization. The World Health report 2002. Midwifery. 2003; 19: 72–73. PMID: 12691085

18. WHO | The world health report 2004—changing history. In: WHO [Internet]. [cited 22 Nov 2016]. http:// www.who.int/whr/2004/en/

19. Wang Y, Wang L, Qu W. New national data show alarming increase in obesity and noncommunicable chronic diseases in China. Eur J Clin Nutr. 2016;

20. Molarius A, Lindén-Boström M, Granström F, Karlsson J. Obesity continues to increase in the majority of the population in mid-Sweden-a 12-year follow-up. Eur J Public Health. 2016; 26: 622–627. doi: 10. 1093/eurpub/ckw042 PMID: 27074794

21. Harris MI, Flegal KM, Cowie CC, Eberhardt MS, Goldstein DE, Little RR, et al. Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. The Third National Health and Nutrition Examination Survey, 1988–1994. Diabetes Care. 1998; 21: 518–524. PMID: 9571335

22. Harris MI, Hadden WC, Knowler WC, Bennett PH. Prevalence of diabetes and impaired glucose toler- ance and plasma glucose levels in U.S. population aged 20–74 yr. Diabetes. 1987; 36: 523–534. PMID: 3817306

23. Haley MJ, Lawrence CB. Obesity and stroke: Can we translate from rodents to patients? J Cereb Blood Flow Metab Off J Int Soc Cereb Blood Flow Metab. 2016;

24. Guo Y, Yue X-J, Li H-H, Song Z-X, Yan H-Q, Zhang P, et al. Overweight and Obesity in Young Adult- hood and the Risk of Stroke: a Meta-analysis. J Stroke Cerebrovasc Dis Off J Natl Stroke Assoc. 2016;

25. González-Gómez FJ, Pérez-Torre P, De-Felipe A, Vera R, Matute C, Cruz-Culebras A, et al. Stroke in young adults: Incidence rate, risk factors, treatment and prognosis. Rev Clínica Esp Engl Ed. 2016; 216: 345–351.

26. Mitchell AB, Cole JW, McArdle PF, Cheng Y-C, Ryan KA, Sparks MJ, et al. Obesity increases risk of ischemic stroke in young adults. Stroke. 2015; 46: 1690–1692. doi: 10.1161/STROKEAHA.115.008940 PMID: 25944320

27. McAuley PA, Blair SN. Obesity paradoxes. J Sports Sci. 2011; 29: 773–782. doi: 10.1080/02640414. 2011.553965 PMID: 21416445

28. Dehlendorff C, Andersen KK, Olsen TS. Body mass index and death by stroke: no obesity paradox. JAMA Neurol. 2014; 71: 978–984. doi: 10.1001/jamaneurol.2014.1017 PMID: 24886975

29. Ramos JM, Padilla S, Masiá M, Gutiérrez F. A bibliometric analysis of tuberculosis research indexed in PubMed, 1997–2006. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2008; 12: 1461–1468.

30. Centers for Disease Control (CDC). A strategic plan for the elimination of tuberculosis in the United States. MMWR Morb Mortal Wkly Rep. 1989; 38: 269–272. PMID: 2495428

31. Frick M. Flatlined: US government investments in tuberculosis research and development, 2009–2012. Treat Action Group. http://www.treatmentactiongroup.org/tbrd2014/usg

32. Starke JR, Cruz AT. The Global Nature of Childhood Tuberculosis. PEDIATRICS. 2014; 133: e725–e727. doi: 10.1542/peds.2013-4139 PMID: 24515521

33. Fitchett JR, Head MG, Atun R. Infectious disease research investments follow colonial ties: question-able ethics. Int Health. 2014; 6: 74–76. doi: 10.1093/inthealth/iht036 PMID: 24464047

34. Morel CM. Neglected diseases: under-funded research and inadequate health interventions. EMBO Rep. 2003; 4: S35–S38. doi: 10.1038/sj.embor.embor851 PMID: 12789404

35. Yamey G. The world's most neglected diseases. BMJ. 2002; 325: 176–177. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1123710/ PMID: 12142292

36. Marsh K. Malaria-a neglected disease? Parasitology. 1992; 104: S53–S69. PMID: 1589300

37. The Global Fund. In: The Global Fund to fight AIDS, Tuberculosis and Malaria [Internet]. [cited 23 Nov 2016]. http://www.theglobalfund.org/en/

38. Pigott DM, Atun R, Moyes CL, Hay SI, Gething PW. Funding for malaria control 2006–2010: A compre-hensive global assessment. Malar J. 2012; 11: 246. doi: 10.1186/1475-2875-11-246 PMID: 22839432

39. Akachi Y, Atun R. Effect of Investment in Malaria Control on Child Mortality in Sub-Saharan Africa in 2002–2008. PLOS ONE. 2011; 6: e21309. doi: 10.1371/journal.pone.0021309 PMID: 21738633

40. Pérez-Gómez B, Martínez C, Navarro C, Franch P, Galceran J, Marcos-Gragera R, et al. The moderate decrease in invasive cervical cancer incidence rates in Spain (1980–2004): limited success of opportu-nistic screening? Ann Oncol Off J Eur Soc Med Oncol. 2010; 21 Suppl 3: iii61–68.

41. Lyrdal D, Aldenborg F, Holmberg E, Peeker R, Lundstam S. Kidney cancer in Sweden: a decrease in incidence and tumour stage, 1979–2001. Scand J Urol. 2013; 47: 302–310. doi: 10.3109/00365599.2012.732608 PMID: 23137102

42. Murphy CC, Sandler RS, Sanoff HK, Yang YC, Lund JL, Baron JA. Decrease in Incidence of Colorectal Cancer Among Individuals 50 Years or Older After Recommendations for Population-based Screening. Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc. 2016;

43. Jacobsen R, Møller H, Thoresen SØ, Pukkala E, Kjaer SK, Johansen C. Trends in testicular cancer inci-dence in the Nordic countries, focusing on the recent decrease in Denmark. Int J Androl. 2006; 29: 199–204. doi: 10.1111/j.1365-2605.2005.00605.x PMID: 16371112

44. Giles GG, Bell R, Farrugia H, Thursfield V. Decrease in breast cancer incidence following a rapid fall in use of hormone replacement therapy in Australia. Med J Aust. 2009; 190: 164; author reply 164–165. PMID: 19203322

45. Hou N, Huo D. A trend analysis of breast cancer incidence rates in the United States from 2000 to 2009 shows a recent increase. Breast Cancer Res Treat. 2013; 138: 633–641. doi: 10.1007/s10549-013-2434-0 PMID: 23446808

46. Crocetti E, Ciatto S, Buzzoni C, Zappa M. Prostate cancer incidence rates have started to decrease in central Italy. J Med Screen. 2010; 17: 50–51. doi: 10.1258/jms.2009.009123 PMID: 20356946

47. Annertz K, Anderson H, Palmér K, Wennerberg J. The increase in incidence of cancer of the tongue in the Nordic countries continues into the twenty-first century. Acta Otolaryngol (Stockh). 2012; 132: 552–557.

48. Aschebrook-Kilfoy B, Schechter RB, Shih Y-CT, Kaplan EL, Chiu BC-H, Angelos P, et al. The clinical and economic burden of a sustained increase in thyroid cancer incidence. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2013; 22: 1252–1259.

49. Pandeya N, McLeod DS, Balasubramaniam K, Baade PD, Youl PH, Bain CJ, et al. Increasing thyroid cancer incidence in Queensland, Australia 1982–2008—true increase or overdiagnosis? Clin Endocri-nol (Oxf). 2015;

50. Polednak AP. Increase in Distant Stage Breast Cancer Incidence Rates in US Women Aged 25–49 Years, 2000–2011: The Stage Migration Hypothesis. J Cancer Epidemiol. 2015; 2015: 710106. doi: 10.1155/2015/710106 PMID: 25649489

51. Dikshit RP, Yeole BB, Nagrani R, Dhillon P, Badwe R, Bray F. Increase in breast cancer incidence among older women in Mumbai: 30-year trends and predictions to 2025. Cancer Epidemiol. 2012; 36: e215–220. doi: 10.1016/j.canep.2012.03.009 PMID: 22521561

52. Youlden DR, Cramb SM, Dunn NAM, Muller JM, Pyke CM, Baade PD. The descriptive epidemiology of female breast cancer: An international comparison of screening, incidence, survival and mortality. Cancer Epidemiol. 2012; 36: 237–248. doi: 10.1016/j.canep.2012.02.007 PMID: 22459198

53. Gompel A, Plu-Bureau G. Is the decrease in breast cancer incidence related to a decrease in postmenopausal hormone therapy? Ann N Y Acad Sci. 2010; 1205: 268–276. doi: 10.1111/j.1749-6632.2010.05664.x PMID: 20840283

54. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature. 1976; 260: 170–173. PMID: 176594

55. Weiss SR, Varmus HE, Bishop JM. Cell-free translation of purified avian sarcoma virus src mRNA. Virology. 1981; 110: 476–478. PMID: 6261453

56. Czernilofsky AP, Levinson AD, Varmus HE, Bishop JM, Tischer E, Goodman HM. Nucleotide sequence of an avian sarcoma virus oncogene (src) and proposed amino acid sequence for gene product. Nature. 1980; 287: 198–203. PMID: 6253794

57. Jenkins JR, Rudge K, Chumakov P, Currie GA. The cellular oncogene p53 can be activated by mutagenesis. Nature. 1985; 317: 816–818. PMID: 3903515

58. Le Beau MM, Westbrook CA, Diaz MO, Rowley JD, Oren M. Translocation of the p53 gene in t(15;17) in acute promyelocytic leukaemia. Nature. 1985; 316: 826–828. PMID: 3929142

59. Szabo CI, King MC. Inherited breast and ovarian cancer. Hum Mol Genet. 1995; 4 Spec No: 1811–1817.

60. Rebbeck TR, Friebel TM, Mitra N, Wan F, Chen S, Andrulis IL, et al. Inheritance of deleterious mutations at both BRCA1 and BRCA2 in an international sample of 32,295 women. Breast Cancer Res BCR. 2016; 18: 112. doi: 10.1186/s13058-016-0768-3 PMID: 27836010

61. Anderson JL, Horne BD, Stevens SM, Grove AS, Barton S, Nicholas ZP, et al. Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation. 2007; 116: 2563–2570. doi: 10.1161/CIRCULATIONAHA.107.737312 PMID: 17989110

62. Choices NHS. Predictive genetic tests for cancer risk genes—NHS Choices [Internet]. 25 Nov 2016 [cited 28 Nov 2016]. http://www.nhs.uk/conditions/predictive-genetic-tests-cancer/pages/introduction.aspx

63. Abbott A. US mental-health chief: psychiatry must get serious about mathematics. Nat News. 2016; 539: 18.

64. Scott JG, Fletcher AG, Anderson ARA, Maini PK. Spatial Metrics of Tumour Vascular Organisation Predict Radiation Efficacy in a Computational Model. PLoS Comput Biol. 2016; 12: e1004712. doi: 10.1371/journal.pcbi.1004712 PMID: 26800503

65. Figueredo GP, Siebers P-O, Owen MR, Reps J, Aickelin U. Comparing stochastic differential equations and agent-based modelling and simulation for early-stage cancer. PloS One. 2014; 9: e95150. doi: 10.1371/journal.pone.0095150 PMID: 24752131

66. Chaplain MAJ, Anderson ARA. Mathematical modelling, simulation and prediction of tumour-induced angiogenesis. Invasion Metastasis. 1996; 16: 222–234. PMID: 9311387

67. Plank MJ, Sleeman BD, Jones PF. A mathematical model of tumour angiogenesis, regulated by vascular endothelial growth factor and the angiopoietins. J Theor Biol. 2004; 229: 435–454. doi: 10.1016/j.jtbi.2004.04.012 PMID: 15246783

68. Evans CJ, Phillips RM, Jones PF, Loadman PM, Sleeman BD, Twelves CJ, et al. A mathematical model of doxorubicin penetration through multicellular layers. J Theor Biol. 2009; 257: 598–608. doi: 10.1016/j.jtbi.2008.11.031 PMID: 19183560

69. Groh CM, Hubbard ME, Jones PF, Loadman PM, Periasamy N, Sleeman BD, et al. Mathematical and computational models of drug transport in tumours. J R Soc Interface. 2014; 11: 20131173. doi: 10.1098/rsif.2013.1173 PMID: 24621814

70. Crowell HL, MacLean AL, Stumpf MPH. Feedback mechanisms control coexistence in a stem cell model of acute myeloid leukaemia. J Theor Biol. 2016; 401: 43–53. doi: 10.1016/j.jtbi.2016.04.002 PMID: 27130539

71. MacLean AL, Filippi S, Stumpf MPH. The ecology in the hematopoietic stem cell niche determines the clinical outcome in chronic myeloid leukemia. Proc Natl Acad Sci U S A. 2014; 111: 3883–3888. doi: 10.1073/pnas.1317072111 PMID: 24567385

72. Powathil GG, Swat M, Chaplain MAJ. Systems oncology: towards patient-specific treatment regimes informed by multiscale mathematical modelling. Semin Cancer Biol. 2015; 30: 13–20. doi: 10.1016/j.semcancer.2014.02.003 PMID: 24607841

73. Barbolosi D, Ciccolini J, Lacarelle B, Barlési F, André N. Computational oncology—mathematical modelling of drug regimens for precision medicine. Nat Rev Clin Oncol. 2016; 13: 242–254. doi: 10.1038/nrclinonc.2015.204 PMID: 26598946

74.   Kershaw SK, Byrne HM, Gavaghan DJ, Osborne JM. Colorectal cancer through simulation and experiment. IET Syst Biol. 2013; 7: 57–73. PMID: 24046975

75.   Byrne HM. Dissecting cancer through mathematics: from the cell to the animal model. Nat Rev Cancer. 2010; 10: 221–230. doi: 10.1038/nrc2808 PMID: 20179714

76.   Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981; 290: 457–465. PMID: 7219534

77.   Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988; 239: 487–491. PMID: 2448875

78.   Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983; 306: 234–238. PMID: 6316146

79.   Gupta A, Shridhar K, Dhillon PK. A review of breast cancer awareness among women in India: Cancer literate or awareness deficit? Eur J Cancer Oxf Engl 1990. 2015; 51: 2058–2066.

80.   Kyle RG, Macmillan I, Rauchhaus P, O'Carroll R, Neal RD, Forbat L, et al. Adolescent Cancer Education (ACE) to increase adolescent and parent cancer awareness and communication: study protocol for a cluster randomised controlled trial. Trials. 2013; 14: 286. doi: 10.1186/1745-6215-14-286 PMID: 24011093

81.   Grimmett C, Macherianakis A, Rendell H, George H, Kaplan G, Kilgour G, et al. Talking about cancer with confidence: evaluation of cancer awareness training for community-based health workers. Perspect Public Health. 2014; 134: 268–275. doi: 10.1177/1757913914534840 PMID: 25169613

82.   Sanz J, Fayad ZA. Imaging of atherosclerotic cardiovascular disease. Nature. 2008; 451: 953–957. doi: 10.1038/nature06803 PMID: 18288186

83.   Mukherjee S. The Emperor of All Maladies [Internet]. Simon and Schuster; 2011. http://www.simonandschuster.com:80/books/The-Emperor-of-All-Maladies/Siddhartha-Mukherjee/9781439170915

84.   van Weert H. "The emperor of all maladies": Towards an evidence-based integrated cancer survivorship care in general practice. Eur J Gen Pract. 2016; 22: 69–70. doi: 10.1080/13814788.2016.1180361 PMID: 27292291