# Can literature analysis identify innovation drivers in drug discovery?

*Pankaj Agarwal and David B. Searls*

Abstract | Drug discovery must be guided not only by medical need and commercial potential, but also by the areas in which new science is creating therapeutic opportunities, such as target identification and the understanding of disease mechanisms. To systematically identify such areas of high scientific activity, we use bibliometrics and related data-mining methods to analyse over half a terabyte of data, including PubMed abstracts, literature citation data and patent filings. These analyses reveal trends in scientific activity related to disease studied at varying levels, down to individual genes and pathways, and provide methods to monitor areas in which scientific advances are likely to create new therapeutic opportunities.

*Computational Biology Department, GlaxoSmithKline Pharmaceuticals, 709 Swedeland Road, PO BOX 1539, King of Prussia, Pennsylvania 19406, USA.*
*Correspondence to P.A. e-mail: Pankaj.Agarwal@gsk.com*
doi:10.1038/nrd2973

Much has been written on the subject of managing innovation in the pharmaceutical industry[1–4]. However, given that the discovery of new drugs often arises from foundational academic research[5–7], the challenge could be reframed as one of effectively recognizing proximal drivers of innovation. It would be advantageous to the drug discovery process to be able to systematically identify therapeutic areas, specific diseases or pathways in which basic scientific understanding is increasing rapidly, and in a manner that is likely to enable new interventions. This could guide the long-term investment that is necessary to build capabilities in those areas, and possibly help in recognizing near-term opportunities. Such assessments are typically made in a subjective and intuitive fashion and, although nothing can replace professional vigilance and the wisdom of experience, expert judgments might benefit from objective metrics and models[8]. Even marginal improvements could be highly beneficial when applied over large drug discovery portfolios.

Unmet medical need and commercial potential, which are considered the traditional drivers of drug discovery, may be seen as providing ‘pull’ for pharmaceutical discovery efforts. Yet, if a disease area offers no ‘push’ in the form of new scientific opportunities, no amount of pull will lead to new drugs — at least not mechanistically novel ones. The trend among payers and health-care providers to require such novelty suggests greater weight should be placed on push in determining investment, and it is becoming clear that even purely commercial considerations favour innovation[4]. Moreover, in the post-genomic era, the identification of areas of rapid scientific advance will grow in importance as the emphasis shifts from target identification to a deeper understanding of targets in the full disease context[6].

## Measuring push

Various factors can influence scientific innovation in the context of drug discovery, and several of these can be tracked over time (FIG. 1). An obvious contributor to the development of science is public investment — as reflected, for example, in the annual budget of the US National Institutes of Health (NIH), which covers internal expenditures and external grants, and therefore a large proportion of US biomedical-research funding. Another factor is the available pool of scientists, to measure which we can tally the number of doctorates awarded by US universities in biology and chemistry, as tracked by the US National Science Foundation (the National Science Foundation Science and Engineering Statistics; see Further information). Publication may be seen as an overall measure of scientific activity (BOX 1), and this can be established by counting the number of articles retrieved from the PubMed database under various selection criteria, such as year of publication. Finally, the ultimate metric of interest is the number of novel marketed drugs, which can be assessed in terms of annual approvals of new molecular entities (NMEs) by the US Food and Drug Administration (FDA). These measures are by no means perfect (being, for example, highly United States-centric), but they have the advantage that reasonably consistent data are publicly available over many decades (FIG. 1), whereas most other trend analyses

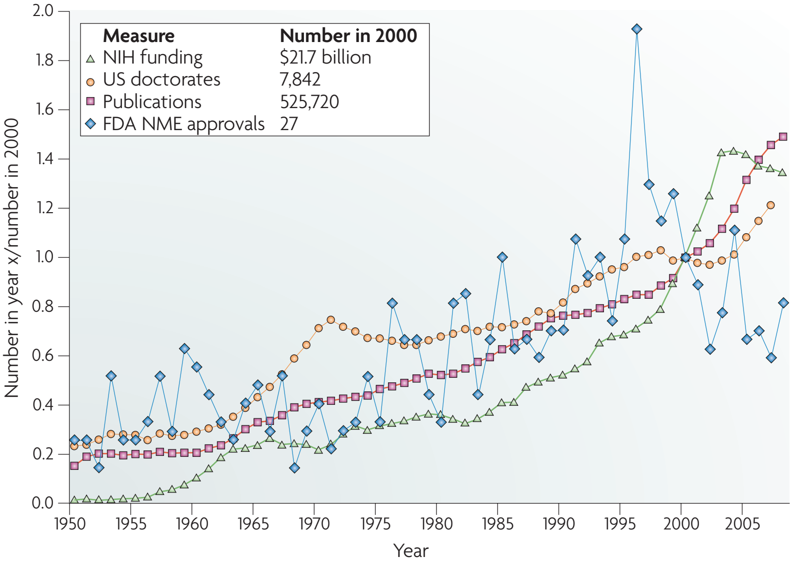


Figure 1 | **NIH funding, US biology and chemistry doctorates awarded, PubMed publications and FDA NME approvals by year.** Data are presented as multiples of the values in the year 2000, which are given in the inset key. National Institutes of Health (NIH) funding data (green triangles) were taken from REF. 54 and adjusted for inflation. Doctorates awarded in the US in either biology or chemistry (orange circles) were taken from REFS 55,56. Publications data (red squares) were determined by counts of articles returned by PubMed when restricted to individual years by the standard filter function for publication date. US Food and Drug Administration (FDA) approvals of new molecular entities (NMEs) (blue diamonds) were determined using the Drug Approval Reports form on the US FDA Center for Drug Evaluation and Research website (see Further information). The number of NME approvals was determined from the tables based on the 'New Drug Application Chemical Type' column entries. When multiple approvals of the same compound in different formulations occured in the same month, these were counted only once. We excluded technetium-based imaging reagents, for which 15 NME approvals were granted for diagnostic kits[7].

are conducted over much shorter time frames and often use proprietary or estimated data — for example, on overall industry investment[9].

In this broad perspective, it is not surprising to see a general upward trend in all measures, albeit with notable variations. It is instructive to examine these variations when considering the relationships between the measures. For example, like the rest of US government investment in science and engineering in the late 1950s, NIH funding increased sharply at this time. In the following decade there was a rapid growth in the number of doctorates awarded, which peaked sharply around 1971 and then subsided into nearly two decades of stasis. However, the period of most rapid growth in the number of Ph.D.s awarded corresponded to constant NIH budgets (adjusted for inflation), illustrating the difficulty of evaluating cause and effect among such measures. Whether the increases in funding and Ph.D.s arose from the same direct stimulus, or the availability of funding influenced the awarding of Ph.D.s, it is not surprising that there should be several years' lag in the number of Ph.D.s awarded relative to funding, given the time required for a doctoral education. Such a hypothesis might account for the evident 'overshoot' in the number of Ph.D.s awarded. However, it is also possible that external confounding variables contributed to one or both trends, including factors such as inflation (reflecting the general business environment), competing spending on new government health initiatives at the time — such as Medicare and Medicaid — and even military conflict, which may have influenced graduate school enrolment and research budgets[10].

The number of Ph.D.s awarded returned to a nearly constant level in recent years, while NIH funding underwent its steepest increases, and then increased again when NIH budgets began to decline. Again, this may be due in part to a natural time lag between these two variables. In addition, some careers may have been more attractive during the economic boom times of the 'dot-com bubble', and others more attractive after the bubble burst. Once again, it is apparent that interpreting such metrics is complicated by the likelihood of time lags in any relationship between them, as well as by the confounding effects of external events that are not directly observed; both of these phenomena are familiar in the practice of time series analysis, as performed by engineers, economists and others[11].

Publication rates also began an upward trend around 1960, but this trend has continued monotonically and with less short-term variation to the present day. The number of publications doubled during the 1960s, but this does not seem to have fully reflected either the sharp increase or the subsequent decline in the number of Ph.D.s awarded during this time. However, if it is assumed that the publication of results from a scientific project follow several years after its original funding, the curves for NIH funding and Pubmed counts seem to track each other — although again, publication rates seem to be buffered against the short-term vagaries of funding.

Much has been made of the recent downward trend in the number of FDA approvals, which began in the mid to late 1990s. This has famously run counter to sharp increases in pharmaceutical R&D investment[9] and public funding, illustrated in FIG. 1 (with the exception of the recent decline in public funding). Although the decrease in NME approvals does coincide with a period in which Ph.D. production was constant, the lag times involved — for example, in bringing a drug from a foundational scientific discovery to the point of approval — would seem to preclude an immediate relationship of FDA approvals to any of the variables in FIG. 1. Rather, a number of explanations have been proposed, ranging from basic flaws in R&D strategies such as over-reliance on genomics, to the introduction of more stringent standards of drug approval bodies[12].

Although there are likely to be confounding external influences, we note the following points regarding the long-term trend. First, the smaller number of data points for FDA approvals makes it inevitable that there will be greater statistical variance in this measure than in the others observed here. If, for example, the spike in approvals in 1996 is discarded as an outlier, the general upward trend can be viewed as continuing at least until the year 2000. Second, until 2000, there were at least three decades during which NME approvals seem

### Box 1 | Bibliometrics and scientometrics

The field of bibliometrics, by which the quantity and character of scientific publications are studied, has long been a mainstay of attempts to measure scientific progress. As a set of methodologies, this field has been driven primarily by practitioners of library science and the sociology of science. The more general field of scientometrics, which uses bibliometrics as a tool to study science in terms of growth and interaction patterns, social structures, costs and other parameters, was pioneered by the historian of science Derek de Solla Price. He first called attention to the exponential growth of scientific literature, based on publication counts[47], and went on to consider the network properties of citations among papers[48].

Citation analysis was greatly advanced by Eugene Garfield, who devised the impact factor — an important and often controversial tool for evaluating the influence of publications and, directly or indirectly, the careers of academics[49,50]. Bibliometric analyses have been used in various applications involving narrowly specified subject areas, including fields relevant to biotechnology[17]. Garfield was the first to use science citation data in a highly directed fashion, in a study that analysed the growth of the field of apoptosis research[51]. Soon after that, it was proposed that medical subject headings (MeSH) from the Medline database could be used to effectively determine the impact factors of topics, as opposed to journals or authors[52]. Bibliometrics has also been applied to patent filings, as in the work of Murray tracing the "network of patents, papers, inventors and authors" in the field of tissue engineering[37], and in the work of Ducor more generally relating co-authorship and co-inventorship[3]. These studies are important forerunners of the work presented in this Analysis.

to track with the general upward trends in both public funding and publication, especially with consideration for appropriate time lags. Third, during the 1960s there was another apparent extended decline in FDA approvals, albeit with a considerably lower baseline, which preceded the sustained increase in productivity for the remainder of the century. (This is often attributed to the thalidomide tragedy which, although largely avoided in the United States, led to much stricter FDA regulations — for example, requiring demonstration of efficacy and not just safety for the first time in 1962 (REF. 13).) Taken together, these observations reinforce the point that short-term trends can be deceptive.

The push for drug discovery — that is, the generation of new and innovative science that can be expected to lead to novel therapeutics — is not easily reduced to a simple metric as there are numerous interdependent factors, latencies and external influences that are unpredictable and unquantifiable. However, for the remainder of this paper we will adopt publication rates as the best available surrogate. Publication is generally nearer to the event of actual discovery than either the funding of research or the training of scientists involved. Its growth seems to be steadier and less immediately susceptible to business cycles and other confounders than that of other variables, while still tracking reasonably closely to new drug approvals over the last half of the twentieth century, given appropriate lag times. There are enough data on publication rates to give sufficient sample sizes for statistical analysis. Perhaps most importantly, they are easily classified by topic and so can be analysed by therapeutic area or even more specific categories. Although publication is not a perfect indicator of innovation, as the acknowledged repository of record for scientific discovery, it is not likely to be improved upon for the purpose of this Analysis.

## Measuring pull

We now turn to the question of measuring the traditional forms of pull for pharmaceutical enterprises. To approximate the pull of medical need and commercial potential, the World Health Organization (WHO) Global Burden of Disease survey[14] provides useful information. For the reference year 2002, it estimated straightforward mortality from various causes and disease burden to society, as measured by days of life lost to either premature death or disability, weighted by severity (the World Health Report; see Further information). We can directly compare these surrogates for push and pull by plotting publication rates in various therapeutic areas, as determined by appropriate PubMed searches (see Supplementary information S1 (table)), against the relative disease burden associated with those therapeutic areas (that is, the burden of a disease relative to the total burden of all diseases) (FIG. 2). We suggest that global disease burden can be primarily considered as a measure of medical need, whereas narrowing the focus to the developed world places greater emphasis on commercial potential. There is a positive correlation between global disease burden and rates of publication of scientific articles, when analysed by therapeutic areas used by the WHO (FIG. 2) ($R^2 = 0.37$, $p < 0.01$; that is, using linear regression the probability that the variables are independent of one another is less than 0.01, and the proportion of the variation in publication rate that can be accounted for by a correlation with relative disease burden is 37%). The correlation is greater in the data for the developed world ($\mathrm{R}^2 = 0.72$), which differs from the global data primarily in a considerably reduced impact of infectious and parasitic diseases, respiratory infections and perinatal conditions (which include factors related to infant mortality).

It is not surprising that scientific output correlates more strongly with the pull of disease burden in a population that is better able to afford scientific research, although it is encouraging that infectious and parasitic diseases have high publication rates given their relatively lower disease burden in the developed world compared with the developing world. Perinatal conditions, by contrast, have low publication rates. The paucity of publications in this area might be attributable to a lack of scientific push: these conditions are closely associated with poverty (and therefore nutrition and access to basic health care) and there are few scientific mysteries surrounding their cause or cure. Publication rates may be reflective not only of push based on areas of rapid scientific advance, but also of pull based on areas most favoured by public funding and private investment. The focus on these areas is driven by medical need and commercial prospects, potentially creating positive feedback loops[15]. Public funding is based on promising and rigorously reviewed science, but is also subject to policy that is driven by medical need; this is evident in the response to the HIV epidemic. However, HIV and AIDS research funding has led to important scientific advances and has enabled new scientific opportunities[16].

The notions of push and pull that we have introduced are unlikely to be completely independent of one another and separable, given the overlap of their causal factors.

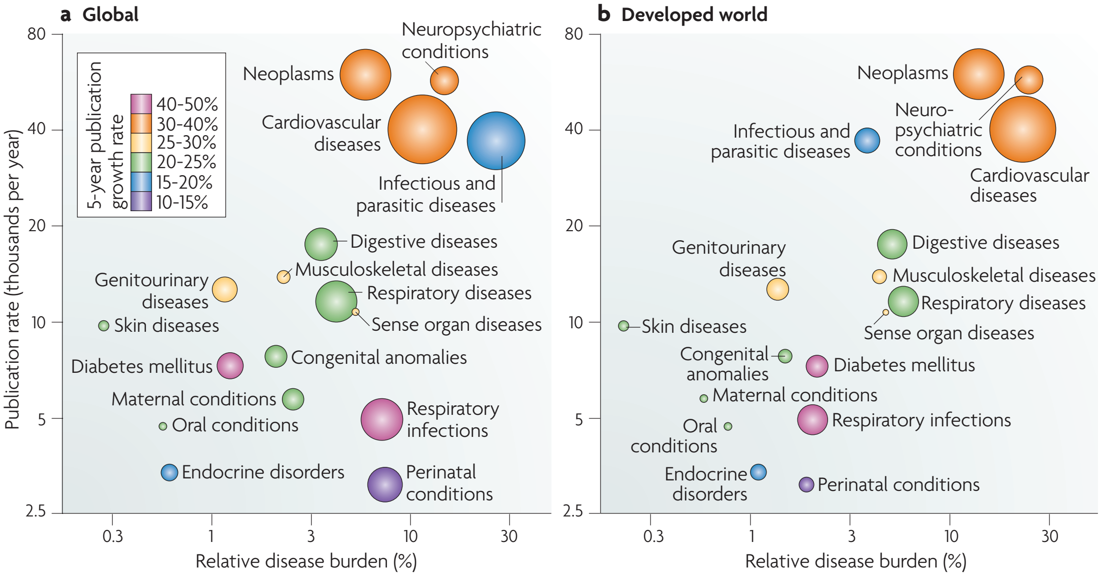


Figure 2 | **Rate of scientific publication versus relative disease burden for key therapeutic areas.** Horizontal axes represent the relative burdens of various disease categories to society, as determined by the World Health Organization (WHO) in 2002. The disability-adjusted life year (DALY) metric for each disease area, which combines a population's years of life lost to each cause with the years of life lived with disabilities (weighted by severity), was divided by the total DALY for all causes to give the relative disease burdens. The vertical axis indicates the average publication rate in thousands of scientific articles per year, over the years 1998–2007 inclusive. Both axes are on logarithmic scales. Each circle represents a therapeutic area as defined by the WHO disease categories, which were mapped to the US National Library of Medicine's medical subject headings (MeSH) that are used for PubMed queries (see Supplementary information S1 (table)). Part **a** shows overall global figures, whereas part **b** is restricted to the developed world, again as defined by the WHO report. The areas of the circles reflect the annual number of deaths for each cause, as a proportion of the relevant population. Circle radii are therefore scaled to the cube root of mortality rates (corresponding to the cross-section of a sphere), and standardized in each part to cardiovascular deaths, which were maximal in both cases: 16,733,160 globally and 6,333,713 in the developed world. For comparison, deaths from diabetes were 987,816 and 244,599, respectively.

Nevertheless, we think that publication is a useful indicator of trends in scientific innovation that may culminate in new drugs. When data on NME productivity are plotted against pharmaceutical R&D expenditure, there is a clear inverse correlation over the past decade. Although pharmaceutical investment is crucial for drug development and its results contribute to the scientific literature, it is reasonable to assume that commercial investment has been more heavily influenced by the pull of commercial potential than has public investment, and that both are more influenced by unmet medical need than the general scientific literature. However, there remains a substantial fraction of publications that are driven primarily by scientific novelty, whether the work has been explicitly identified as basic research or has arisen through the natural opportunism of the research process (or in the extreme case, by serendipity). It is in the character of the literature to follow the most novel and exciting scientific trends, creating an effect that resembles a chain reaction — for example, as seen with the recent upsurge in the number of papers on microRNA[17].

Moreover, pharmaceutical investment is skewed to late-stage drug development, whereas public funding and publication have a greater impact on scientific novelty related to targets and mechanisms. A 2005 study showed that, of the NMEs that were approved by the FDA from 1998 to 2003, 14% of compounds were directly attributable to university invention (and 28% to biotechnology company R&D)[7]. It is likely that an even larger proportion of target identification and validation has benefited from public research than these figures suggest and should therefore be evident in the scientific literature.

## Publication trends

One particular challenge to analysing publication trends is that long-term averages of publication rates may reflect institutional inertia, whereby established journals, funding bodies, academic careers and infrastructure combine to create a self-perpetuating publication focus[15]. When medium-term rates of change in publication patterns are examined (FIG. 2), there is less correlation with disease burden, but suggestions that changes in publication rates in a given area may also be influenced by scientific opportunity. For example, research in infectious diseases, despite starting at a high rate of publication, is currently one of the slowest growing categories, perhaps reflecting the notorious recent paucity of new antibiotic classes[18–20]. The fastest-growing category is diabetes, which has experienced a late surge in novel target opportunities, second only to oncology[21–23]. However, even for rapidly changing growth trends in publication rates, there may be influences from newly arising medical need; for example, the rapid 5-year growth in publication rate in the area of respiratory infections in 2003 was partly due to the public-health challenges of avian and human influenza and severe acute respiratory syndrome (SARS), and has slowly subsided[24].

To study publication rates by individual therapeutic area, we make use of the medical subject headings (MeSH) terminology — which is used by annotators of PubMed — beginning with the top-level diseases category set (but including also 'mental disorders' from the psychiatry and psychology set, and excluding 'animal diseases' and 'pathological conditions, signs and symptoms'). This categorization differs from that used by the WHO as well as other disease ontologies in common use[25–28], but has the advantage of having been uniformly applied over many years by PubMed annotators to indicate whether a given paper has a given category as a major topic (as opposed to straightforward text searches that may produce incidental occurrences). Technical details of our methods are presented elsewhere[29].

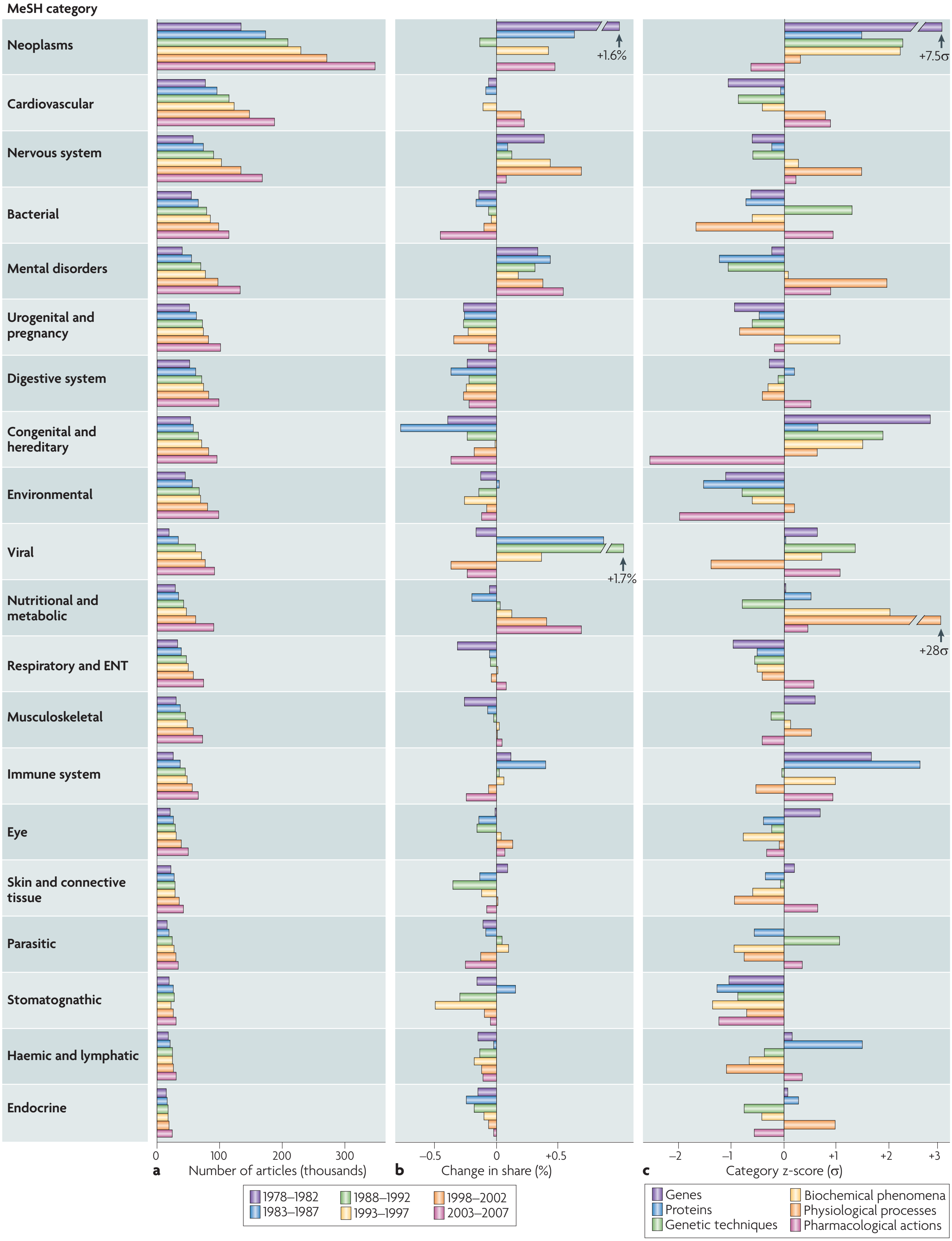
The MeSH hierarchy overlaps itself in important ways (for example, diabetes mellitus is counted under both endocrine system diseases and nutritional and metabolic diseases). Even if the hierarchy itself were non-overlapping, many articles covering several related topics would be classified under more than one heading, as detailed indexing is performed. Therefore, when the entire PubMed corpus is analysed, certain disease headings have a substantial percentage of papers in common. Specifically, 15% of disease area pairings exceed 20% overlap in articles, and several exceed this overlap considerably (see Supplementary information S2 (table)). We mitigate double counting within those pairs by various expedients (see Supplementary information S3 (table)), which generally involves either combining headings with major overlaps or subtracting counts in one category from another so as to assign the 'hit' to only one of the categories. Together, the adjustments made are sufficient to ensure that no pair of disease areas has more than 20% of articles in common.

As an example, one global adjustment was the subtraction of hits for neoplasms from all other categories, as tumours occur in so many tissues and systems. Note that this does not preclude any mention of cancer in an article that is categorized in another disease area, but only the annotation of cancer as a major topic of the article, on a par with the other category; in cases in which the article was annotated with neoplasms as a major category, the article is counted solely under neoplasms in our analysis. Using these rationalized MeSH categories to count articles published over the past 30 years shows that publications in each disease area increased continuously over 5-year periods (FIG. 3a), with the exception of stomatognathic (oral) diseases, on which publications declined in the mid 1990s. The absolute increases are most evident for the areas with the greatest numbers of articles published, such as cancer, cardiovascular research and neurology. In this and all subsequent figures, the order of presentation of MeSH categories is based on decreasing overall 30-year publication counts.

To make differences and trends more apparent, the publication counts can be divided by the total number of articles for all disease areas in each time period, for an indication of the 'market share' of a disease area. These shares vary from 17.2% for cancer to 4.5% for viral disease to 1.4% for endocrine disorders, for the overall 30-year counts. The share in each 5-year period is then compared with the share in the preceding 5 years to show the percentage change in the share of total publications of each disease area (FIG. 3b). It is evident that publications on cancer have increased strongly but unevenly in share, starting from a large initial increase over 1973–1977, which was probably a prolonged effect of the enhanced funding throughout the 1970s associated the US National Cancer Act of 1971 (REF. 30). The other most marked change is for viral disease, the publication counts for which grew sharply in the 1980s, driven by the emergence of HIV and AIDS[31,32], and have since subsided noticeably. Publications on mental disorders have shown the steadiest growth over three decades, and there are upturns in publications on neurology in the late 1990s, and nutritional and metabolic disorders (which, as noted above, include diabetes) in the past decade. Certain other disease areas exhibit a nearly continuous decline in share, including urogenital, digestive, and congenital and hereditary diseases, although again it should be noted that these trends are relative to publications in other areas and not absolute trends. In a few cases the trends may also relate to changes in disease area MeSH annotations over time.

The long-term signals detected by the analyses above generally accord with recognizable trends and events, such as disease emergences and policy shifts. However, shorter-term tendencies are of greater interest for the insights they may provide for current decision making. To characterize such variations, we perform similar analyses on annual counts, rather than 5-year groupings, over the same 30-year period. Although the year-on-year changes in numbers of articles published are too variable to readily illuminate trends, it is possible to visualize rates of change over a 3-year window in a 'heat' map (FIG. 4). This confirms impressions from the more coarse-grained year-on-year analysis, such as the extended expansion in virus research in the 1980s, beginning with the first demonstrations of retroviral involvement in human disease, and reinforced by the identification of HIV as the causative agent for AIDS in the mid 1980s[31]. Between the first incidence of AIDS and the identification of HIV as the causative agent for AIDS, there is also a burst of activity in the immune system category. The activity in stomatognathic diseases in the early–mid 1980s was spread over several topics, including temporomandibular joint disorder and periodontal disease, but this growth is relative to a low baseline rate.

Among more recent short-term trends detected is the surge of activity in the nutritional and metabolic category, as discussed above. A strong increase in the nervous system category just before 2000 may reflect the culmination of the 'Decade of the Brain' (a US initiative from 1990 to 2000, the stated aim of which was "to enhance public awareness of the benefits to be


MeSH category
Neoplasms
Cardiovascular
Nervous system
Bacterial
Mental disorders
Urogenital and pregnancy
Digestive system
Congenital and hereditary
Environmental
Viral
Nutritional and metabolic
Respiratory and ENT
Musculoskeletal
Immune system
Eye
Skin and connective tissue
Parasitic
Stomatognathic
Haemic and lymphatic
Endocrine
+1.6%
+1.7%
+7.5σ
+28σ
0
100
200
300
a
Number of articles (thousands)
−0.5
0
+0.5
b
Change in share (%)
−2
−1
0
+1
+2
+3
c
Category z-score (σ)
1978–1982
1983–1987
1988–1992
1993–1997
1998–2002
2003–2007
Genes
Proteins
Genetic techniques
Biochemical phenomena
Physiological processes
Pharmacological actions

◂ Figure 3 | **Scientific publication over three decades classified by MeSH disease headings.** The disease areas for each column were derived from top-level medical subject headings (MeSH), as described in the main text. **a** | Numbers of articles published by 5-year spans, given in the inset key (left). **b** | Changes in 'market share' of each disease area relative to that of the previous period. Share was calculated by dividing the number of articles in the given disease area by the total number of articles published in all disease areas. Time spans are as in part **a**. Two values that are off-scale are indicated by arrows. **c** | Characterization of disease areas by various additional MeSH categories, indicated in the inset key (right) and described in the main text. The fraction of articles in each disease area with the given annotation is calculated, followed by the mean and standard deviation over all disease areas. The difference between each fraction and the mean is divided by the standard deviation to determine a z-score. In calculating the statistical parameters, two outliers (marked by arrows) were omitted so as not to compress the scale due to a large standard deviation. ENT, ear, nose and throat.

derived from brain research") and especially the impact of enhanced functional imaging capabilities. Both this increase in nervous system research and the expansion in research on mental disorders in the past decade were enabled by strong growth in neurosciences funding from 1995 to 2005 by industry and government[33]. The burst of publication activity in the respiratory category is attributable to SARS, avian influenza and chronic obstructive pulmonary disease. Note that the overlap between respiratory and viral diseases did not reach the threshold requiring separation in our rationalization of MeSH categories, so that these viral infections also occur in the respiratory category.

The visualization of annual publication growth rates (FIG. 4) also makes evident a general decline across many diseases in the early 1990s, which is less apparent in the overall publication data (FIG. 1). Although this reduction in disease-related publication rates ends shortly before the start of the downturn in pharmaceutical NME productivity, it is not possible to infer causality given the complexity of the relationships between the variables, as discussed above.

## Characterizing disease areas

As well as determining the number of articles that are published in a given disease area, additional information about those articles can be assessed using various MeSH annotations that relate to other aspects of content. We compiled several such subclassifications for the years 2003–2007 within each therapeutic area. For example, if an article is annotated with 'genes' or 'proteins' as major headings, it suggests that the research being reported uses a molecular biological approach that may offer specific target opportunities. Such a conclusion is also suggested by the MeSH category 'genetic techniques', referring to the use of any of a wide range of molecular biological techniques, from cloning and sequencing to expression profiling and proteomics. The heading 'biochemical phenomena' encompasses cellular biochemistry and signal transduction, whereas 'physiological actions' includes cell physiology, electrophysiology, homeostasis and similar phenomena. In short, when present on an article with a specific disease area as a major heading, these annotations suggest that the article addresses the molecular basis of a disease and the characterization of potential targets. The heading 'pharmacological actions' suggests the availability of chemical tools or potential drugs. To present these characterizations in a fashion that allows for their comparison, we calculate the fraction of articles in each disease area with a given annotation, and then determine a z-score — that is, the number of standard deviations this fraction is from the mean value for this annotation over all disease areas (FIG. 3c). The histogram shows that cancer has by far the most articles on genes, which is due to a preponderance of oncogenes and specific oncogenic mutations, followed by the congenital and hereditary disease category, which is driven by single-gene Mendelian disorders, and then the immune system category. The congenital and hereditary disease area is also noteworthy for its use of genetic techniques, not surprisingly, although cancer and various infectious diseases are also positive for this subcategory annotation. The immune system category and the haemic and lymphatic disease category are heavily annotated with proteins, due to the prevalence of cell surface and extracellular signalling proteins, immunoglobulins and biopharmaceutical products. As expected, biochemical phenomena and physiological processes feature prominently in the nutritional and metabolic disease area, and biochemical phenomena are also notable in cancer and congenital and hereditary diseases. Other notable hits for physiological processes are neurological and mental disorders, and to a lesser extent urogenital, endocrine and cardiovascular disease. Mental disorders and cardiovascular disease areas arguably exhibit a 'pregenomic' profile: above average in physiological processes and pharmacological actions but below average in the attributes that relate to molecular targets.

It is also possible to examine the past literature on an area of disease using additional MeSH categories selected *ad hoc*. For example, in the field of oncology there have been small but discernible waves of publication arising from new areas of research — including oncogenes (which were first mentioned in more than 1% of cancer articles in 1984, and peaked as a percentage in 1993), apoptosis (beginning in 1995 and peaking in 2004), and gene expression profiling (beginning in 2002 and peaking in 2005) — which have contributed to the overall body of cancer research. Underlying these were a steadier background of growth from topics such as protein kinases (beginning in 1988) and signal transduction (beginning in 1999), which continue to increase up to the present day. The MeSH hierarchy branches out through several levels of classification to some 25,000 subcategories, such that we are able to compile counts for every possible annotation and search for recent changes in publication rates that achieve statistical significance, some of which are described below.

## High-impact publications

The MeSH filters on the vast numbers of articles published may be used to identify the types of publications that provide the greatest scientific push, at least as regards understanding of publication trends at a molecular level. However, the quality of the publications is also important. The numbers of publications cannot have increased

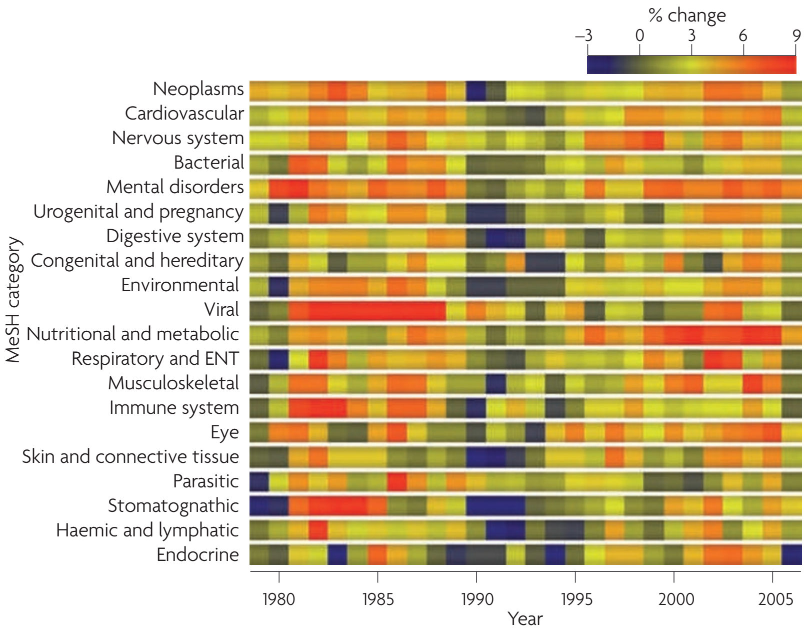


Figure 4 | **Rates of change in scientific publication by year for MeSH disease categories.** The rate of change in the numbers of publications in each disease area (measured as in Figure 2) for each year from 1979 to 2006. Red 'hot spots' indicate surges of publication and dark blue areas represent periods of reduced activity. ENT, ear, nose and throat; MeSH, medical subject heading.

at the rates observed without an approximately corresponding increase in the numbers of journals, and there is no guarantee that this mutual expansion is in proportion to the rate of generation of novel, important, well-executed science. Although restricting attention to a few high-impact journals has the disadvantage of decreasing sample size and therefore statistical power, it has two salutary effects: first, it establishes a 'container' of constant size over time, for which space is at a premium and competition fierce; and second, it increases the chances of capturing papers that are more than ordinarily important, based on editorial standards of scientific impact and broad interest.

For these purposes we used an often-cited ranking of scientific journals based on traditional impact factors combined with the Google PageRank algorithm, which takes into account the prestige of the citing journal[34]. From this list we chose the top three medical journals (*The New England Journal of Medicine, The Journal of the American Medical Association* and *The Lancet*) and the top three non-medical journals (*Nature, Science* and *Cell*). Searches for each disease area were restricted to these journals over the same time periods as in our previous analyses, and radar plots were used to provide multiple dimensions of the data in a visually tractable manner (FIG. 5). In this way, variations among journals in their representation of different disease areas are immediately evident from the shapes of the polygons, and changes over time are apparent from the widths of the coloured bands. Axes are scaled such that the areas of the polygons approximate the fraction of all disease-related articles in each journal that are represented by that disease area. The superimposed dotted hexagons represent the average number of publications on that disease area across all journals from 1978 to 2007.

As in most of the analyses that we have done, publications on cancer dominate, showing strength in all six journals but a marked over-representation in *Cell* and a slight under-representation in the medical journals. The virus category publications show strong growth (for reasons discussed above) in *Science* and *Nature*, and slight over-representation in medical journals. The under-representation of infectious and neurological and psychiatric disease area publications in *Cell* relative to the other journals highlights two caveats to this approach: first, that the stated scopes of journals create differences in the types of submissions the journal attracts, which may vary over time (such as the historic focus of *Cell* on mammalian systems); and second, that major journals tend to spin off new specialty journals (such as *Neuron*) which may be equally competitive but more narrowly targeted, such that impact factors are lower. Indeed, many specialty journals are equally or more likely to publish groundbreaking papers, but the principle of the method we have used is to sample a restricted space at the top of the hierarchy, where the most basic scientific or medical advances are likely to be witnessed. Moreover, altering the analysis to account for these caveats would also introduce ascertainment bias, for reasons such as that most of the journals in question have not existed for the 30-year time span sampled.

The numbers of publications in the cardiovascular research area is considerably higher in the medical journals than in the scientific journals. This publication skew towards medical journals is also a feature of urogenital, digestive, haemic and lymphatic, and endocrine diseases, perhaps reflecting imbalances in the basic versus clinical science that underpins the various therapeutic areas. Publications in the respiratory disease area, however, have grown considerably in *Nature* and *Science* in the most recent 5-year period, much of it due to the interest of these journals in influenza and SARS. We also note recent growth in the number of *Cell* articles on disorders of environmental origin, which is based on small numbers of articles for that disease area and overwhelmingly concerns the topic of DNA damage.

It is also possible to measure article impact independently of journal impact by directly consulting citation data of individual publications, especially given an apparent trend in the scientific community towards citing fewer and more recent articles[35]. For articles published in each year from 2002 to 2006, we determined the top 1% of the most-cited papers that were associated with any disease area, and then determined the relative representation of each disease area in that elite set as a whole (FIG. 6). Each square in the graph represents one of the years measured, from smallest (2002) to largest (2006). Squares shaded red represent those measurements that are significantly greater than the 1% that would be expected by chance ($p < 0.05$ after correction for multiple testing). Similarly, squares shaded blue represent measurements that are significantly below the chance expectation for top-cited articles. Disease

Impact factor
The average number of citations made in a given year to articles that were published in the journal in question during the two preceding years. Impact factors (and many variants) are used to rank journals for their impact on a field, and are sometimes misused in evaluating individual researchers.

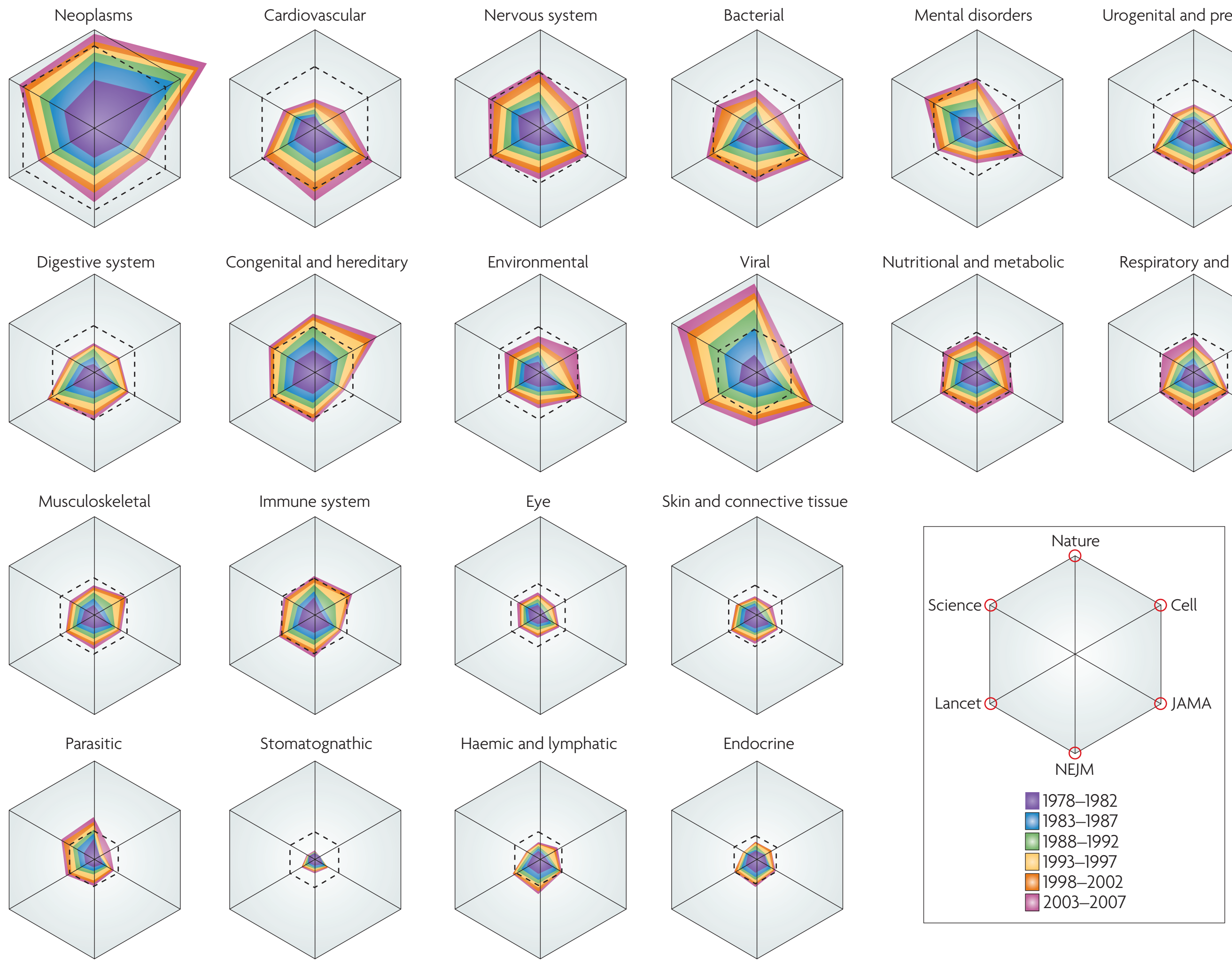


Figure 5 | **Representation of MeSH disease categories in high-impact journals.** Each 'radar plot' indicates the relative extent of representation of the indicated medical subject heading (MeSH) in six journals: *Nature*, *Science*, *Cell*, *The Journal of the American Medical Association* (JAMA), *The New England Journal of Medicine* (NEJM) and *The Lancet*. Journals are oriented around the hexagons according to the key (bottom right), with basic-science journals at the top and medical journals at the bottom. Values on each axis represent the proportion of the given disease area relative to the overall number of disease-related articles in that journal. The outer edge of each hexagon frame represents 25% of disease-related articles, and the axes are scaled by the square roots of the fractions so that the areas of the polygons approximate the overall counts. The nested irregular polygons represent cumulative contributions of successive 5-year spans, as indicated in the key, so that widths of colour bands reflect the relative contributions for those periods. The superimposed dotted hexagons show the percentage represented by that disease area in the entire disease-related scientific literature for the overall period. ENT, ear, nose and throat.

areas that exhibit a strong recent citation record include nutritional and metabolic disorders, cancer and cardiovascular disease, and publications on viral diseases show strong citations in several recent years.

Another measurement of scientific push, and one more closely related to commercial exploitation of scientific discoveries than publication rates, is the rate at which articles lead to intellectual property in the form of patents. We scanned full-text patents from US, European and World Intellectual Property Organization filings and, based on author names, associated those patents to specific journal publications and thence to disease areas (BOX 2). By automating techniques used by others for patent–paper pairing[36,37], we were able to analyse a substantially greater number of patents than was previously possible. Additional integrity checks were applied, such as the proximity of dates to one another. Although most patents cannot be unambiguously associated with specific papers, a sampling suggests that our algorithm is sufficiently accurate to provide a reliable signal. Overall, 3–4% of papers could be associated to patent filings. The same scheme is used for plotting the results for patents as was used for citations (FIG. 6). Again, cancer is found to be the leader in the field, with only the immune system category and recently the virus category also demonstrating greater than average numbers of patent filings. The

### Box 2 | Methods

Publication counts were generated using a freely available tool set of the National Center for Biotechnology Information, eUtils[53]. The PubMed identifiers for each selected disease area, journal and other MeSH categories were downloaded using eUtils. The entire PubMed database was also downloaded and information on publication year was extracted. These data were used to generate counts for each disease area and range of years. The queries corresponding to disease areas were defined in a way that minimized overlap between disease areas (see Supplementary information S2 (table) and S3 (table)). We also examined each disease with restrictions on the publication type to journal, editorial, letter, news and comments, but discovered no unexpected trends. Gene names were identified in abstracts using synonyms from the Human Genome Organization (HUGO), EntrezGene and UniProt databases[29].

Citations were accumulated from all publications until mid 2008. Science Citation Index (SCI) data were obtained from Thomson Reuters. 7.3 million articles from PubMed were mapped to SCI using article titles and journal names, of which 6.3 million were unique matches. This enabled the attachment of citation counts to PubMed articles indexed with MeSH terms. The citation count for a given paper is the cumulative number of publications until mid 2008 that cited the paper.

Published patents were downloaded from Micropatents (Thomson Reuters). This included US, European and World Intellectual Property Organization filings with a primary or secondary classification code that suggested relevance to drug discovery (namely, A61K, A61P, C01, C07, C08, C12N, C12P and C12Q). 134,887 articles that were published in 2002–2006 were mapped to individual patents based on matching subsets of author names. Patent matches to publications were scored by summing the inverse author frequency — which is inversely proportional to the number of patents on which a given author is named as an inventor — of all matching authors on a patent and a publication. The score threshold for a match was reached when at least two authors (with no more than 10 patents each) were named on both the patent and the publication. The publication had to be no longer than 3 years after, and not before, the patent filing year. A sampling of the resulting matches was curated. We estimated ~75% precision, assessed solely on the correspondence of the titles of patents and publications. Although this cannot accurately associate a specific patent with a publication, we think it provides a reasonable sampling across a disease area. Similar patent to publication matching algorithms have been used previously[37].

correlation between number of citations and number of patent filings within disease areas is strong, except for cardiovascular, nutritional and metabolic, and to a lesser extent mental disorders, which show greater strength in the number of citations than in the number of patent filings. A correspondence between the number of citations and the number of patents for individual scientists has been noted before[37], but we observe that this connection may extend to entire disease areas.

## Diseases, genes and pathways

So far we have examined various bibliometric measures at the level of disease areas, restricting our analysis to the top-level MeSH disease categories. This may be appropriate in assessing major organizational commitments to invest in therapeutic areas, but a more frequent need would be to examine specific diseases or other more fine-grained divisions of therapeutic areas for their current level of scientific activity. A narrower focus, whether by time, category or filters of any kind (such as for papers with genes as major topics), will reduce article counts and statistical power. This has the disadvantage of producing a more variable readout that requires careful assessment, but has the great advantage of removing any averaging effect of large categories and allowing small but important signals to emerge. Subdividing disease areas is also more likely to avoid the effects of institutional inertia discussed above, as redirection of resources and attention within a broad research area, which can be detected by examining disease subcategories, is easier than changing career directions and editorial remits to entirely new fields.

We typically assess publication growth in two time frames: the number of publications in the most recent 2 years relative to the 2 years before that, and the number of publications in the most recent 5 years relative to the 5 years before that. This provides short-term trends — 2 years being about the minimum time to accumulate sufficient counts for statistical significance — and medium-term baseline trends for comparison, so as to establish whether a trend is accelerating, decelerating or reversing. We apply the two-sided Fisher's Exact test, with Bonferroni correction for multiple testing, on the large number of MeSH disease categories and subcategories to determine which exhibit recent changes that are significantly different from variation that is expected by chance. MeSH terms that were introduced or have changed their organizational history in the past 10 years are excluded from this analysis as it is difficult to derive accurate quantitative data on them. These excluded diseases are included in Supplementary information S4 (table) as the fact that their organization in the hierarchy has changed may make them intrinsically interesting. Results are only used for diseases that can be clearly defined, rather than broad categories of disease. If a category and one of its immediate subcategories are both found to show significant changes, only one — generally the more specific one — is selected for display.

The disease categories with the most significant recent trends are displayed in a scatterplot with 5-year growth on the horizontal axis and 2-year growth on the vertical axis (FIG. 7a). With respect to a diagonal that represents equal 2-year and 5-year growth rates, publication rates for those diseases above the diagonal are accelerating relative to the 5-year baseline trend, and those for diseases below the diagonal are decelerating. The sizes of the circles indicate the numbers of articles that were published in the past 2 years, and the colours indicate the magnitude of the statistical significance. Insulin resistance shows the strongest growth in publications on both axes, confirming that research on insulin resistance was the main component of recent publication growth in the nutritional and metabolic disease area. Articles on orthomyxoviridae infections exhibit strong expansion beyond the baseline growth, reflecting work on avian influenza. At the other extreme, categories such as hyperlipidaemia and helicobacter infections have shown recent declines in publications. Most 2-year publication trends are reasonably consistent with 5-year trends, with the exception of publications on neoplasm invasiveness, which show a marked recent decline despite extensive medium-term growth.

The same technique can be applied to other annotations besides those given by MeSH categories, provided there is a facility to extract additional features from the PubMed abstracts. For example, to assess growth rates in publications mentioning specific genes, we conduct a search for recognizable gene names (and their synonyms)

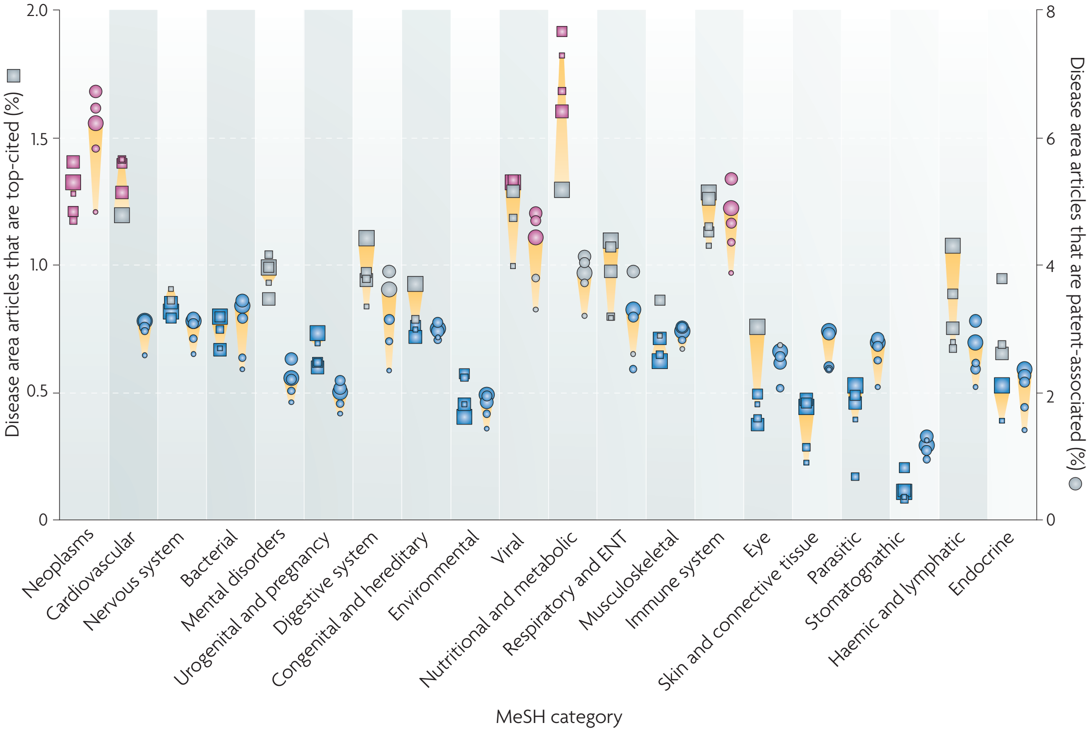


Figure 6 | **Highly-cited articles and associations with patent filings by disease area over 5 years.** This figure shows percentages of highly-cited articles (squares) and patent filings (circles) in each disease area. In each column the icons, from smallest to largest, represent the years 2002–2006. Red icons indicate the values are greater than expected by chance, blue icons represent values that are less than expected by chance and grey icons represent values that are the same as expected by chance, according to a two-sided Fisher's Exact test ($p < 0.05$) with Bonferroni correction for multiple hypothesis testing. Yellow trapezoids connect the icons representing the first and last years of the span, giving some indications of spread and trend, but not of statistical significance. Squares on the left side of each column represent the numbers of highly-cited articles, determined by finding the 1% most-cited articles published in each year in all disease areas and then dividing the contribution of each disease area by this total. Therefore, the expected value for sets of articles chosen at random would be 1%. Circles on the right side of each column represent the percentages of articles in each disease area that are associated with patent filings, as described in BOX 2. ENT, ear, nose and throat; MeSH, medical subject heading.

by methods briefly described in BOX 2 and detailed elsewhere[29] (FIG. 7b). We focus on genes that show growth in publication rates. As might be expected, we find more variability in trends for genes than for disease subcategories, reflected in wider dispersion from the diagonal. So, publications that mention forkhead box P3 (FOXP3) in their abstract have continuously increased in number, reflecting the recent appreciation of the role of this transcription factor in the control of the regulatory T cell lineage in mice[38]. Publications that mention janus kinase 2 (JAK2) exhibit a recent burst in publication rate beyond their consistent medium-term growth, which probably results from the discovery of a variant of this well studied gene that is prevalent in chronic myeloproliferative disorders[39]. By contrast, *ADIPOQ* (adiponectin, C1Q and collagen domain-containing) was only characterized shortly before the medium-term time frame of our analysis. It exhibits strong early publication growth when associated with metabolic syndrome and cardiovascular disease[40] — a growth that has slackened in the near-term time frame.

It is also possible to identify pathways that are associated with the most productive publication record, by taking all publications associated with genes that have shown growth in publication rates in the past 2 years ($p < 0.001$), and mapping them onto manually curated pathways derived from various sources (such as BioCarta). We restrict our attention to those pathways with 10–100 associated genes and identify the genes in each pathway that exhibit strong publication growth. We use a two-tailed Fisher's Exact test to determine which pathways are enriched with such genes to an extent beyond chance expectation. The top pathways, in order of decreasing significance, are toll-like receptor signalling (8 out of 42 genes with significant publication growth), tumour necrosis factor receptor 2 signalling (8 out of 46 genes), interleukin-22 soluble receptor signalling (9 out of 80 genes) and dendritic cells in regulating T helper type 1 ($T_H1$) and $T_H2$ development (5 out of 22 genes) (see Supplementary information S5 (table)).

## Discussion

Bibliometrics provides a set of methods for quantitative analysis of the scientific literature — the repository of knowledge that is important to drug discovery. Although this data collection might seem chaotic, with 'noisy' and

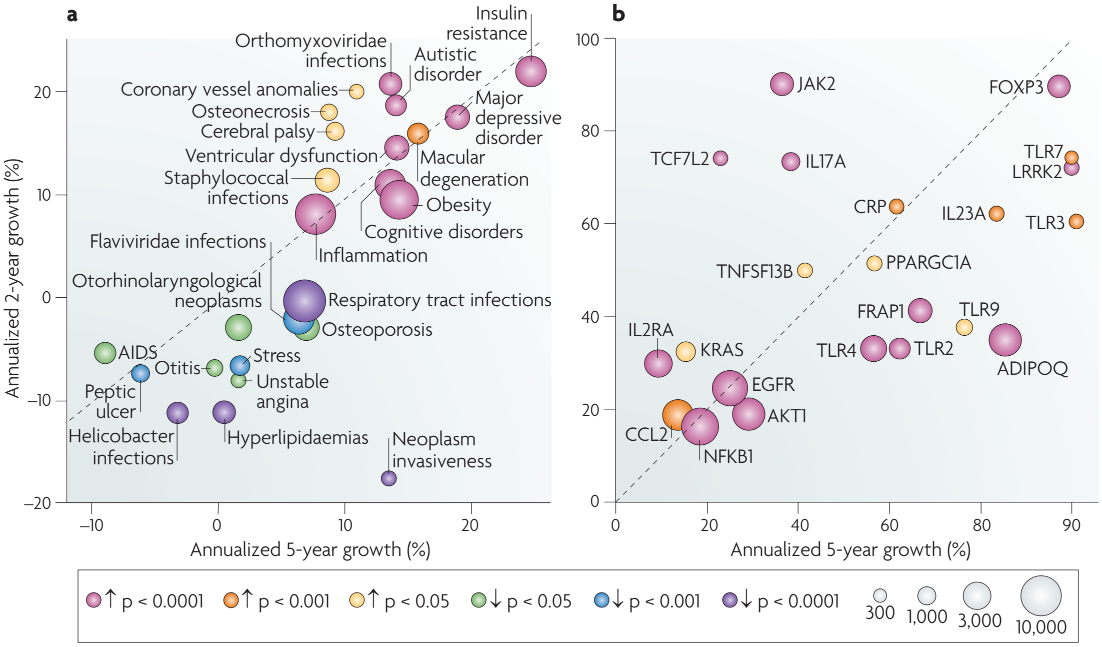


Figure 7 | **Recent growth in publications by disease and by individual gene.** The horizontal axes show percentage growth in 2003–2007 relative to 1997–2002, and the vertical axes show growth in 2006–2007 relative to 2004–2005. The diagonal therefore indicates points for which growth is consistent in the short and medium term; publication rates of points above the diagonal are accelerating and those for points below the diagonal are decelerating. The size of the circles indicates the number of publications in the final 2 years surveyed, as shown in the key (grey circles, scaled to the cube root of publication count). Colours indicate the statistical significance of the 2-year differences after Bonferroni correction for multiple testing, for either positive or negative growth, as defined in the key (coloured circles; arrows pointing up indicate positive growth and arrows  pointing down indicate negative growth). **a** | Changes in publication rates concerning medical subject heading (MeSH) disease subcategories that achieve statistical significance (out of 4,354). Several diseases are not shown, in cases in which their MeSH annotation changed between 1998–2007, or in which the categories were very broad, had fewer than 500 publications in 2006–2007, or were closely related to another disease already shown. Coronavirus infections (mostly due to severe acute respiratory syndrome) were off the scale, with 84% annualized growth on the 5-year axis but 36% annual decline over the past 2 years. **b** | Changes in publication rates concerning genes that achieve statistical significance (out of 13,024). Only genes with over 100 identifiable publications in 2006–2007 are shown. The overall analysis may be biased toward genes with more established and consistent nomenclature, as it was necessary to scan abstracts for gene symbols and known synonyms to derive this data. ADIPOQ, adiponectin, C1Q and collagen domain-containing; CCL2, CC-chemokine ligand 2; CRP, C-reactive protein; EGFR, epidermal growth factor receptor; FOXP3, forkhead box P3; FRAP1, FK506 binding protein 12-rapamycin associated protein 1 (also known as mTOR); IL17A, interleukin-17A; IL-23A, interleukin-23 subunit-α; IL2RA, interleukin-2 receptor-α; JAK2, janus kinase 2; LRRK2, leucine-rich repeat kinase 2; NFKB1, nuclear factor κB1; PPARGC1A, peroxisome proliferator-activated receptor-γ, coactivator 1α; TCF7L2, transcription factor 7-like 2; TLR, toll-like receptor; TNFSF13B, tumour necrosis factor (ligand) superfamily, member 13b.

sometimes confusing signals over time, bibliometrics can be used to monitor long-term tendencies and possibly even highlight medium- or near-term anomalies that can be readily investigated further.

We have suggested that bibliometric analysis can be used to indicate the extent of novel research activity at the level of broad therapeutic areas and at the level of individual diseases, as well as to address additional, specific questions. However, it must be acknowledged that selecting certain aspects of the data can compromise statistical analysis: smaller sample sizes reduce the power of any tests applied, and seemingly paradoxical statistical effects may arise when data are divided into subsets[41]. These will be limiting factors in one's ability to focus attention on scientific ideas of high current interest, and perhaps this is the point at which human instinct and judgment are needed.

The scientific literature is not homogeneous, and subsets of the data may be uneven in various respects. Trends that are seen when more specific selection criteria are applied to data, such as individual diseases or genes for which there are limited numbers of publications, will be more susceptible to effects such as investigator bias, which tends to emphasize the more eminent and/or well funded workers in the field. Different therapeutic areas are likely to vary in their publication rates and practices, based on differing cultures among scientific disciplines, the number of available journals (possibly with different remits and customs) and other factors. With regard to citation analysis, it has been noted that citation practices depend on the field of study and that, although these can be normalized, the results may not be consistent[42]. However, this can be seen as resulting from different, legitimate citation practices, rather than a lack of robustness[42]. Moreover, such differences among fields can often be normalized by examining rates of change rather than absolute differences.

The literature is also heterogeneous (which, as we have indicated, may be partly addressed by focusing on higher-quality journals and measures of impact) and varies in type, ranging from standard research reports to review articles to letters to the editor. Although it might be argued that reviews and letters do not add to the store of scientific knowledge and should not be included in a measure of scientific activity, it could also be argued that reviews tend to be written on topics of current interest, and that letters may be especially indicative of intense scientific activity.

Many of these hurdles are likely to be overcome by averaging data and focusing on rates of change. The fundamental question remains the validity of scientific publication as a metric of innovation. As noted at the beginning of this Analysis, the push for drug discovery comprises a complex of interdependent factors that contribute to the advance of science, and interact in feedback loops with the pull of unmet medical need and commercial potential. Between these and the end product of proven novel therapies, which necessarily lag behind the initial scientific discoveries by a number of years, the publication record, which is more amenable to detailed analysis, could constitute a fitting surrogate for innovation.

However, it may not be possible to prove this, at least at the level of individual targets. Some years before the development of recent novel drugs, one might expect to see flurries of publications relating to their targets or pathways. However, advances in the use of genomic technologies for target discovery will have been a factor only in the development of more recent drugs, introducing a potential bias into analyses of this field. Moreover, even some recent drugs did not originate from the identification of a molecular target. For example, although the first thiazolidinedione was marketed in 1997 for type 2 diabetes, peroxisome proliferator-activated receptor-γ was only recognized as its target 2 years before that, when the drug class had long been in development[43]. The trend towards repurposing drugs means that a burst of new research in a disease area may lead to the novel application of a drug long after it is first marketed — for example, the recent interest in thalidomide for its immunomodulatory properties that has produced a sharp increase in publication after decades of dormancy[44].

Therefore, the use of bibliometric methods to indicate potentially fruitful areas for drug discovery relies on the success of modern target identification and validation techniques, and on the assumption that increased levels of scientific activity, indicated by publication rates, will point to specific targets, as well as pathways, systems and focused disease areas, that are suitable for novel therapies. For the identification of new therapeutic opportunities, we are currently investigating the predictive power of bibliometric data combined with various other data sources, such as industry pipeline activity. Others have used sophisticated literature-mining and natural language processing techniques that could extract a more detailed knowledge than that afforded by the MeSH annotation we have primarily employed[29]. Information flows in biological communities have also been recently investigated through citation analysis[45]. In summary, what we have previously referred to as the bibliome seems to offer many additional approaches to enhancing decision-making in drug discovery[46].

1. Zhong, X. & Moseley, G. B. Mission possible: managing innovation in drug discovery. *Nature Biotech.* **25**, 945–946 (2007).
2. Ullman, F. & Boutellier, R. A case study of lean drug discovery: from project driven research to innovation studios and process factories. *Drug Discov. Today* **13**, 543–550 (2008).
3. Sams-Dodd, F. Optimizing the discovery organization for innovation. *Drug Discov. Today* **10**, 1049–1056 (2005).
4. Cohen, F. J. Macro trends in pharmaceutical innovation. *Nature Rev. Drug Discov.* **4**, 78–84 (2005).
5. Chin-Dusting, J., Mizrahi, J., Jennings, G. & Fitzgerald, D. Finding improved medicines: the role of academic–industrial collaboration. *Nature Rev. Drug Discov.* **4**, 891–897 (2005).
6. Vallance, P. & Levick, M. Drug discovery and development in the age of molecular medicine. *Clin. Pharmacol. Ther.* **82**, 363–366 (2007).
7. Kneller, R. The origins of new drugs. *Nature Biotech.* **23**, 529–530 (2005).
8. Davenport, T. H. & Harris, J. G. *Competing on Analytics: The New Science of Winning.* (Harvard Business School Press, Boston, Massachusetts, 2007).
9. US Department of Health and Human Services. Innovation or stagnation? Challenge and opportunity on the critical path to new medical products. *The National Institute for Pharmaceutical Technology and Education website* [online] <http://www.nipte.org/docs/Critical_Path.pdf> (2004).
10. Card, D. & Lemieux, T. Going to college to avoid the draft: the unintended legacy of the Vietnam war. *Am. Econ. Rev.* **91**, 97–102 (2001).
11. Shumway, R. H. & Stoffer, D. S. *Time Series Analysis and Its Applications.* (Springer, New York, 2005).
12. Ruffalo, R. R. Why has R&D productivity declined in the pharmaceutical industry? *Expert Opin. Drug Discov.* **1**, 99–102 (2006).
13. Bren, L. Frances Oldham Kelsey: FDA medical reviewer leaves her mark on history. *FDA Consum.* **35**, 24–29 (2001).
14. Mathers, C. D. *et al.* The global burden of disease in 2002: data sources, methods and results. *Global Programme on Evidence for Health Policy*. Discussion Paper No. 54. World Health Organization (2003; revised 2004).
15. Teitelbaum, M. S. Research funding: structural disequilibria in biomedical research. *Science* **321**, 644–645 (2008).
16. Cohen, J. Bang for the buck. *Science* **321**, 518–519 (2008).
17. Taroncher-Oldenburg, G. & Marshall, A. Trends in biotech literature 2006. *Nature Biotechnol.* **25**, 961 (2007).
18. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature Rev. Drug Discov.* **6**, 29–40 (2007).
19. Vicente, M. *et al.* The fallacies of hope: will we discover new antibiotics to combat pathogenic bacteria in time? *FEMS Microbiol. Rev.* **30**, 841–852 (2006).
20. Coates, A. R. & Hu, Y. Novel approaches to developing new antibiotics for bacterial infections. *Br. J. Pharmacol.* **152**, 1147–1154 (2007).
21. Ashiya, M. & Smith, R. E. T. Non-insulin therapies for type 2 diabetes. *Nature Rev. Drug Discov.* **6**, 777–778 (2007).
22. Das, S. K. & Chakrabarti, R. Non-insulin dependent diabetes mellitus: present therapies and new drug targets. *Mini Rev. Med. Chem.* **5**, 1019–1034 (2005).
23. Morral, N. Novel targets and therapeutic strategies for type 2 diabetes. *Trends Endocrinol. Metab.* **14**, 169–175 (2003).
24. Webby, R. J. & Webster, R. G. Are we ready for pandemic influenza? *Science* **302**, 1519–1522 (2003).
25. Caviedes, J. E. & Cimino, J. J. Towards the development of a conceptual distance metric for the UMLS. *J. Biomed. Inform.* **37**, 77–85 (2004).
26. Wang, X. *et al.* Automating terminological networks to link heterogeneous biomedical databases. *Medinfo* **11**, 555–559 (2004).
27. Patel, C. O. & Cimino, J. J. Mining cross-terminology links in the UMLS. *AMIA Annu. Symp. Proc.* **2006**, 624–628 (2006).
28. Pedersen, T., Pakhomov, S. V., Patwardhan, S. & Chute, C. G. Measures of semantic similarity and relatedness in the biomedical domain. *J. Biomed. Inform.* **40**, 288–299 (2007).
29. Agarwal, P. & Searls, D. B. Literature mining in support of drug discovery. *Brief Bioinform.* **9**, 479–492 (2008).
**In this article, the authors of the Analysis provide details of methods used herein and review wider applications of literature mining that are specifically aimed at drug discovery.**
30. Kalberer, J. T. Jr & Newell, G. R. Jr. Funding impact of the National Cancer Act and beyond. *Cancer Res.* **39**, 4274–4284 (1979).
31. Karpas, A. Human retroviruses in leukaemia and AIDS: reflections on their discovery, biology and epidemiology. *Biol. Rev. Camb. Philos. Soc.* **79**, 911–933 (2004).

32. Cohen, J. HIV/AIDS. Where have all the dollars gone? *Science* **321**, 520 (2008).
33. Dorsey, E. R. *et al.* Financial anatomy of neuroscience research. *Ann. Neurol.* **60**, 652–659 (2006).
34. Bollen, J., Rodriquez, M. A. & Van de Sompel, H. Journal status. *Scientometrics* **69**, 669–687 (2006).
35. Evans, J. A. Electronic publication and the narrowing of science and scholarship. *Science* **321**, 395–399 (2008).
**A much discussed study showing that online publishing, and the ease of following hyperlinks, tends to channel researchers towards a narrower and more recent set of publications, with a possible loss of diversity and historical perspective.**
36. Ducor, P. Intellectual property: coauthorship and coinventorship. *Science* **289**, 873–875 (2000).
37. Murray, F. Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. *Res. Policy* **31**, 1389–1403 (2002).
38. Fontenot, J. D. & Rudensky, A. Y. A well adapted regulatory contrivance: regulatory T cell development and the forkhead family transcription factor Foxp3. *Nature Immunol.* **6**, 331–337 (2005).
39. Mesa, R. A. New insights into the pathogenesis and treatment of chronic myeloproliferative disorders. *Curr. Opin. Hematol.* **15**, 121–126 (2008).
40. Gable, D. R., Hurel, S. J. & Humphries, S. E. Adiponectin and its gene variants as risk factors for insulin resistance, the metabolic syndrome and cardiovascular disease. *Atherosclerosis* **188**, 231–244 (2006).
41. Ramanana-Rahary, S., Zitt, M. & Rousseau, R. Aggregation properties of relative impact and other classical indicators: convexity issues and the Yule–Simpson paradox. *Scientometrics* **79**, 311–327 (2009).
**Although somewhat technical, this paper describes important statistical artefacts that can arise when classifications of the scientific literature are aggregated or subdivided, including the reversal of certain trends.**
42. Zitt, M., Ramanana-Rahary, S. & Bassecoulard, E. Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalisation. *Scientometrics* **63**, 373–401 (2005).
43. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptorγ (PPARγ). *J. Biol. Chem.* **270**, 12953–12956 (1995).
44. Calabrese, L. & Fleischer, A. B. Thalidomide: current and potential clinical applications. *Am. J. Med.* **108**, 487–495 (2000).
45. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105**, 1118–1123 (2008).
46. Searls, D. B. Mining the bibliome. *Pharmacogenomics J.* **1**, 88–89 (2001).
47. De Solla Price, D. J. *Little Science, Big Science.* (Yale University, New Haven, 1963).
48. Price, D. J. Networks of scientific papers. *Science* **149**, 510–515 (1965).
**A classical paper by the founder of scientometrics, which showed that networks of citations among scientific papers obey a power law distribution. It was published many decades before the study of such scale-free networks achieved prominence.**
49. Lawrence, P. A. The mismeasurement of science. *Curr. Biol.* **17**, R583–R585 (2007).
50. Lawrence, P. A. The politics of publication. *Nature* **422**, 259–261 (2003).
51. Garfield, E. & Melino, G. The growth of the cell death field: an analysis from the ISI-Science citation index. *Cell Death Differ.* **4**, 352–361 (1997).
**In this paper, the originator of the impact factor, Eugene Garfield, uses bibliometrics to trace and analyse the historical development of the field of apoptosis.**
52. Takahashi, K., Aw, T. C. & Koh, D. An alternative to journal-based impact factors. *Occup. Med. (Lond.)* **49**, 57–59 (1999).
53. Sayers, E. & Wheeler, D. Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils). *The NCBI website* [online] <http://www.ncbi.nlm.nih.gov/bookshelf/picrender.fcgi?book=coursework&part=eutils&blobtype=pdf> (2004).
54. American Association for the Advancement of Science. Historical data on federal R&D, FY 1978–2009. *The AAAS website* [online] <http://www.aaas.org/spp/rd/hist09p2.pdf> (2008).
55. National Science Foundation. Doctoral degress awarded, by detailed field: 1920–99. *The National Science Foundation website* [online] <http://www.nsf.gov/statistics/nsf06319/pdf/tabs1.pdf> (accessed 2009).
56. Falkenheim, J. C. & Fiegener, M. K. 2007 records fifth consecutive annual increase in US doctoral awards. *The National Science Foundation website* [online] <http://www.nsf.gov/statistics/infbrief/nsf09307/nsf09307.pdf> (2008).

## Acknowledgements

We gratefully acknowledge P. Vallance for posing the problem, V. Kumar for implementing the methods to identify disease areas and genes with strong publication growth, and R. Liu for suggesting the use of common authors to map patents to publication, and thereby associate patents with MeSH key words. D. Rajagopalan, K. Kabnick, L. Liu, T. White, W. Reisdorf and Y. Li provided technical assistance and valuable suggestions.

## FURTHER INFORMATION

**David B. Searls's homepage:** http://www.med.upenn.edu/apps/faculty/index.php/g306/c425/p6363

**National Science Foundation Science and Engineering Statistics:** http://www.nsf.gov/statistics/

**US FDA Center for Drug Evaluation and Research:** http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm

**World Health Report:** http://www.who.int/whr/2004/en/

## SUPPLEMENTARY INFORMATION

**See online article:** S1 (table) | S2 (table) | S3 (table) | S4 (table) | S5 (table)

**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**