High-Impact and Transformative Science (HITS) Metrics: Definition, Exemplification, and Comparison

Joseph Staudt[1], Huifeng Yu[2], Robert P. Light[3], Gerald Marschke[2,4,*], and Katy Börner[3,5], Bruce A. Weinberg[1,4]

[1] Ohio State University, Columbus, Ohio USA

[2] University at Albany, State University of New York, Albany, New York, USA

[3] CNS, School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA

[4] National Bureau of Economic Research, Cambridge, Massachusetts, USA

[5] Indiana Network Science Institute, Indiana University, Bloomington, Indiana, USA

*Corresponding author email: marschke@albany.edu

Abstract

Countries, research institutions, and scholars are interested in identifying and promoting high-impact and transformative scientific research. This paper presents novel metrics that use and extend bibliometric and linguistic approaches to identify high-impact and transformative works. The 11 general metrics can be grouped into seven types: Radical-Generative, Radical-Destructive, Risky, Multidisciplinary, Wide Impact, Growing Impact and High Impact. A total of 10,778,696 articles available in Thomson Reuters' Science Citation Index Expanded™ linked to MEDLINE are used to exemplify and validate the metrics. Publication years are grouped into six 5-year periods spanning 1983-2012 and 6,159 comparable MeSH terms that characterize the fields to which each article belongs. The analysis is conducted at the level of a field-period pair, of which 15,051 have articles and are used in this study. Analysis show that transformativeness is positively related to impact (ρ*=.401*), but no evidence that transformative work is adopted slowly or that the generation of important new concepts coincides with the obsolescence of existing concepts.

1. Introduction

Countries, research institutions, and scholars have prioritized high-impact and transformative scientific research. The National Science Board (NSB) argues that while research that has the potential to transform science "is inherently less predictable in its course and eventual outcomes, it is, nonetheless, absolutely essential for our national advancement and for the advancement of science as a whole (NSB, 2007)."

Recognizing the importance of transformative research, the National Institutes of Health (NIH) and National Science Foundation (NSF) both instituted initiatives to support transformative research. Yet, no standard metrics exist to identify transformative research, which limits our ability to answer even basic questions about it. Such metrics are essential if we want to answer even such fundamental questions as: How frequent is transformative research? How important is transformative research for scientific progress? Does the prevalence of transformative research vary over time or across fields? To what extent do high impact and transformative work overlap? How do the demographics (in terms of gender, race, age, national origin) of fields, the structure of scientific networks, or the funding environment affect the production, diffusion, and reception of transformative research?

A National Science Board report from 2007 argues:

> Science progresses in two fundamental and equally valuable ways: The vast majority of scientific understanding advances incrementally, with new projects building upon the results of previous studies or testing long-standing hypotheses and theories. This progress is evolutionary—it extends or shifts prevailing paradigms over time. The vast majority of research conducted in scientific laboratories around the world fuels this form of innovative scientific progress. Less frequently, scientific understanding advances dramatically, through the application of radically different approaches or interpretations that result in the creation of new paradigms or new scientific fields. This progress is revolutionary, for it transforms science by overthrowing entrenched paradigms and generating new ones. The research that comprises this latter form of scientific progress … [is] termed transformative research... (National Science Board, 2007).

This paper develops metrics to identify those scientific fields and periods of time ("field-period pairs") in which high-impact work and transformative work was done. We begin by grounding our work in established conceptualizations of transformative research from NIH, NSB, and NSF. These conceptualizations identify seven aspects of transformative work. Transformative work is seen to be:

radical in the sense of (1) generating important new ideas and (2) making existing ideas obsolete or less salient (3) risky, (4) multidisciplinary, (5) having a broad impact, (6) having an impact that builds over time, and (7) being high impact. We then develop eleven metrics that correspond to these aspects of transformative work using rich characterizations of citation patterns as well as natural language processing techniques. We next use factor analysis to identify the combination of our eleven metrics that best characterize the seven aspects of transformative work. Finally, we reduce the dimensionality of the metrics of transformativeness (other than impact) into a single measure of transformativeness. The various metrics of impact are very closely related to each other. The behavior of our metrics of transformativeness largely correspond to existing conceptualizations, but provide insights. Conventional citation measures of impact are related to transformativeness, but our metrics show substantial independent variations in transformativeness for a given level of impact ($\rho=.401$). Thus, impact and transformativeness are empirically (as well as conceptually) distinct, each representing distinctive, cohesive phenomena.

Measures of scientific output and creativity in the social science literature rarely extend beyond publication counts, perhaps weighted by some journal ranking, and citation counts, which do not adequately distinguish work that is influential within a paradigm from work that is influential and also path-breaking and therefore do not allow separate analysis of impact and transformativeness in science. Recent work has sought to address deficiencies of standard citation methods (e.g., Wang, Song, and Barabasi [2013] and Hutchins, Yuan, Anderson, and Santangelo [2015]). Acemoglu, Akcigit, and Celik [2015] use a range of rich characterizations of citations to identify the most innovative work. Wang, Veugelers, and Stephan (2016) identify novel research from unique combinations of citations. Funk and Owen-Smith (Forthcoming) uses shifts in citation patterns to identify work that consolidates or destabilizes existing technologies. Evans and Foster (2016) overview approaches to identifying novelty and develop a unifying simulation approach.

Cronin and Sugimoto (2014, 2015) review a wide range of scholarly metrics that are commonly used in citation and scholarly impact analysis but also in academic auditing. While traditional metrics use a quantitative analysis of publications, authors, bibliographic references, and related concepts, novel metrics also take into account text, acknowledgments, endorsements, downloads, recommendations, blog posts, and tweets. They argue that multi-dimensional metrics—also called mixed indicators—are most valuable as the performance of a person, institution, or country cannot be adequately measured by any single indicator. This is in line with research by Bollen et al. (2009) which compared 39 existing and proposed metrics of scholarly impact that are calculated on the basis of both citation and usage log data. They performed a principal component analysis of the rankings produced by these metrics to investigate how the different metrics relate to each other, and how accurately and completely they express scientific impact. They too conclude that the notion of scientific impact is a multi-dimensional construct and multiple metrics are needed to cover impact.

Research by Hanning et al. (2011) developed and validated mixed indicators that help identify emerging research areas. Bettencourt, Kaiser, and Kaur (2009) use the evolution of scientific collaboration networks to trace the evolution of fields. The work presented in this paper is novel as it focuses on the development of metrics that support the identification of high-impact and transformative science (HITS).

2. Conceptualization of Transformative Work

Consistent with NSB (2007), scientific works vary continuously along two dimensions: 1) the extent to which they are radical (versus incremental) and 2) their impact, from low to high. These dimensions are illustrated in Figure 1. Most work in science is incremental, increasing knowledge and practices within an established paradigm or theoretical framework. As knowledge, products, and practices accumulate incrementally, moderate amounts of knowledge and practices become obsolete. High-impact incremental work naturally has a large impact on a field but lies within an existing paradigm. Consequently, high-impact, incremental work does not make obsolete a large amount of research (relative to its impact). Radical work differs from incremental work in that it represents a break from an existing paradigm. The highest-impact radical work is transformative and game-changing, fundamentally altering a discipline, making existing theories, paradigms, and knowledge obsolete, or at least less salient. It also generates new research opportunities, potentially across many fields. Of course, not all radical work has a high impact. Low-impact radical work neither contributes to an established paradigm nor successfully replaces one. Our distinction between incremental and radical work parallels Kuhn's (1947) distinction between normal and revolutionary science. We hypothesize that this classification applies to non-scientific innovation and across research motivations (as in Stokes [1997]).

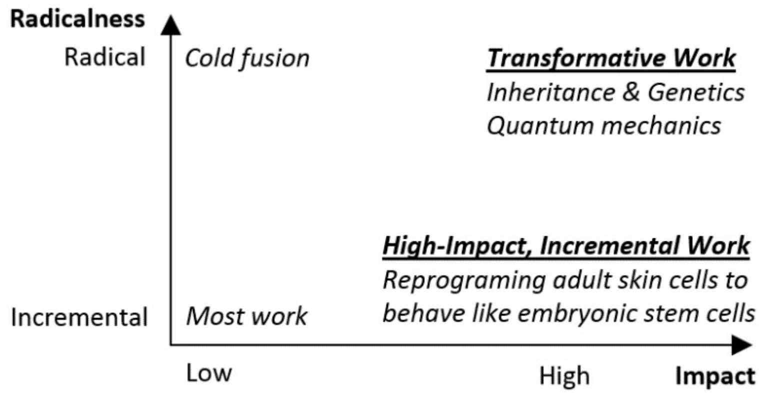

Figure 1. Classification of scientific work by radicalness and impact, with examples.

Figure 1 provides examples that illustrate our classification, although we caution that a rigorous classification requires formal metrics such as those proposed in this paper.

Lower-left quadrant: Most scientific work is incremental and has a comparatively low impact. For example, in genetics and related fields, the discovery that two genes interact to produce a particular phenotype often is a publishable result. Dissecting the molecular mechanism controlling gene expression, however, is a more difficult and significant advance; this is the type of finding that is published in the top journals in molecular biology and genetics, such as *Cell*, and that has a higher impact.

Lower-right quadrant: According to *Time Magazine*, one of the top-ten scientific discoveries of 2007 was a method to reprogram skin cells to behave like embryonic stem cells. This discovery is likely to have a high impact because it provides a means to obtain stem cells without destroying human embryos, thus bypassing legal and ethical roadblocks to stem cell research. From a clinical perspective, it offers the potential of regenerating from a punch of skin a patient's pancreas, liver, muscle, and so forth that are exactly like the patient's own. This method is credited to two research teams working independently, each publishing its findings in November 2007. As of July 2016, according to Google Scholar, these two articles received 7,752and 11,707 citations, respectively. Despite its high impact, this discovery did not radically alter a scientific paradigm. We therefore place this advance high on the impact scale but low on the radicalness scale.

Top-right quadrant: Quantum mechanics is a canonical example of transformative work in the 20th century, as it marked a shift from classical physics, changed physicists' view of the world, and impacted other fields, such as chemistry. Examples of transformative research in biomedicine range from a series of breakthroughs in genetics and inheritance from Mendel's genetic theory, to the discovery of the link between the DNA and inheritance, to the identification of the structure of DNA, which launched the fields of genetics and molecular genetics.

Top-left quadrant: Low-impact radical works are works that fail or lead to dead ends (e.g., cold fusion; see Bettencourt, Kaiser, and Kaur [2009]) and radical works that impact a small area or make a small advance to a paradigm.

Scientific contributions can be classified as "conceptual" (e.g., discovery of the DNA structure) or "technical," involving the development of methods or tools. Our classification of scientific works applies to both. Insofar as a technical contribution incrementally improves existing techniques and does not radically alter practices or overturn the theoretical framework, paradigm, or body of knowledge, it will be incremental. A new tool or method that renders existing tools or methods obsolete or whose application directly changes the theoretical paradigm in use is transformative. The invention of the tunneling microscope was transformative because it enabled new inquiries that ultimately resolved longstanding, fundamental questions and created new bodies of knowledge and even new fields (Chen, 2007). Another example of a transformative scientific discovery of a technical nature is the discovery in 1998 of RNA interference (RNAi), a natural process by which cells silence the activity of specific genes. Prior to the discovery of RNAi, nearly the only method available to disable a gene in mammals was by creating knockout or transgenic animal models, a very time-intensive and uncertain process. RNAi-based gene suppression is now the state-of-the-art method by which scientists can "knock down" specific genes in cells to learn about gene function (Gao and Zhang, 2007).

As indicated, in seeking to develop HITS metrics, we draw on existing conceptualizations from NIH, NSF, and the NSB. In recent years, the NIH has established programs that specifically target transformative research. The objective of NIH's Roadmap Transformative Research Projects Program (R01) is to support "exceptionally innovative and/or unconventional research projects with the potential to create or overturn fundamental paradigms. These projects tend to be inherently risky and may not fare well in conventional NIH review… The primary emphasis of the Transformative Research Award is to support research on bold, paradigm-shifting but untested ideas" (NIH, 2015). The Common Fund's NIH Director's Transformative Research Award is intended to "support research on bold, paradigm-shifting but untested ideas" (NIH, 2015). The NSF defines transformative research as involving "ideas, discoveries, or tools that radically change our understanding of an important existing scientific or engineering concept or educational practice or leads to the creation of a new paradigm or field of science, engineering, or education. Such research challenges current understanding or provides pathways to new frontiers (NSF, 2015)". It describes transformative research as "revolutionizing entire disciplines; creating entirely new fields; or disrupting accepted theories and perspectives—in other words, those endeavors which have the potential to change the way we address challenges in science, engineering, and innovation." Because potentially transformative research challenges the research agendas of experts on review panels, it may not receive a fair hearing. Also as the NSB notes, transformative research frequently crosses disciplinary lines adding to the challenge of evaluating the work. Nonetheless it views transformative research as being "of critical importance in the fast-paced, science and technology-intensive world of the 21st Century (NSB, 2007)" and thus should be of paramount importance in determining how scarce funding is allocated.

These descriptions point to seven aspects of transformative work, many of which appear in multiple conceptualizations, and are often described using the same vocabulary. Given the lack of formal metrics for transformative work, we view these aspects as potentially characterizing transformative work, with the actual features of transformative work being an empirical question that we seek to address in this work. The characteristics and how they map back to the conceptualizations, are outlined in Appendix A, Table A.1, and described in detail below. All metrics for transformative work are computed at the level of field-period pairs as motivated in Section 3. The seven characteristics of transformative work and the metrics we develop to measure them are:

(1) Radical-Generative—Transformative research is viewed as critical because it generates radical new paradigms, theories, perspectives, and fields. We measure the generative aspect of transformative research using the introduction of heavily used new terms, measured by a metric called *Concepts*, and the utilization of important new terms, called *BMentN*, where N indicates the number of years (0, 3, 5, 10, ∞) since the term was first used in an article. (Here and elsewhere, "*B*" and "*F*" prefixes indicate backward and forward measures; "*Ment*" indicates mentions of concepts, "*Cite*" indicates citations; and "*Herf*" and "*Age*" indicate Herfindahl measures of dispersion and ages.)

(2) Radical-Destructive—In creating radical new paradigms, transformative research is seen to render large portions of existing knowledge obsolete (or at least less salient). Backward citations, captured by a metric called *BCiteAge*, indicate the extent to which current research draws on prior work. Jones and Weinberg (2011) show that backward citations ages contract during scientific revolutions.

(3) Risky—Because it represents a substantial departure from existing work, the existing conceptualizations view transformative work as risky. The risky nature is one reason why transformative work might not receive the support that it merits in funding reviews and why it is especially important to be able to identify and support it. One natural measure of risk is the variance in forward citations received by the articles published in a field-period pair, here called *FCiteVar*.[1]

(4) Multidisciplinary—Transformative work is viewed as more likely to draw on knowledge from many fields. We use Herfindahl indices (a standard measure of dispersion used by economists) to measure the breadth of fields that are cited in articles and call this metric *BHerfCite*. In addition, we generate metrics for the breadth of important new terms that the articles in a field-period pair draw on. Specifically, we define *BHerfMentN*, where N indicates the number of years (0, 3, 5, 10, ∞) since the term was first introduced into the literature.

(5) Wide Impact—Just as transformative work is viewed as more likely to draw on a wide range of knowledge, it is seen to be more likely to have a wide impact. We measure the breadth of impact using Herfindahl indices of the range of fields that cite articles (using a metric we call *FHerfCite*) and the range of fields that use the terms introduced by articles (using a metric called *FHerfMent*).

(6) Growing Impact—Because it is radical, the impact of transformative work is seen to take a while to accumulate. We measure the time path of utilization of transformative work using the mean time elapsed between when an article is published and the forward citations it receives and call the corresponding metric *FCiteAge*.

(7) High impact—In order for a radical work to be transformative, it must be high impact, so we view this aspect of transformative work as somewhat definitional. Put differently, works that are as radical as transformative work, but that do not have the same impact will not transform fields. We define the metric *FCiteMean* as the mean forward citation count and the percentiles

[1] In addition to the riskiness of research in a field-period pair, this measure reflects differences in the importance of work done in a field stemming from other sources.

of the distribution of forward citation counts as *FCiteN,* where N indicates the percentile of the citation distribution (25, 50, 75, 90, 95, 99, 99.9, 99.99).

3. Data Acquisition and Preparation

Two datasets are used to construct and exemplify the eleven metrics: 1) MEDLINE® 2014 baseline files distributed by the National Library of Medicine (NLM) containing 22,376,811 articles published between 1809 and 2014 and 2) 15,085,762 articles from Thomson Reuters' Science Citation Index Expanded™ published between 1950 and May 20, 2014, the day our data was acquired. After taking the intersection of the two data sources, we are left with 13,737,835 articles published between 1950 and 2014. See Table 1 for details.

Table 1: Article Counts

Data Source	**Articles**	**With Restrictions**
MEDLINE 2014 Baseline *Published 1809-2014*	22,376,811	20,667,693*
Web of Science *Published 1950-May 20, 2014*	15,085,762	15,080,131**
Intersection		13,737,835
Published 1983-2012		10,778,696

*There are three restrictions on articles in the MEDLINE data: 1) the article must be the first version of an article, 2) the article must have "MEDLINE" status, and 3) the article must be tagged with at least one 4-digit MeSH term. For details on the version and status of MEDLINE articles (NLM, 2016). For details on 4-digit MeSH terms see below description and Appendix C.

**There is one restriction on articles in the Web of Science (WOS) data: A small number of our WOS records map to a PMID to which other WOS records map. We retain the earliest WOS ID that maps to each PMID, reducing our WOS articles by 5,631 or .037% of our 15,085,762 WOS records.

Period Identification

We are interested in generating two sets of metrics—one based on text analysis and another based on citations patterns. Since article abstracts are important for generating our text-based metrics, and MEDLINE's coverage of abstracts is poor before 1980, we limit our sample to articles published in 1983 or later. Since citations take time to accumulate and our data ends in 2014, we limit our sample to articles published in 2012 or earlier. As seen in Table 1, restricting our sample to articles published between 1983 and 2012 leaves us with 10,778,696 articles with which to compute our metrics.

Field Identification

The 10,778,696 articles in our analysis sample are tagged with Medical Subject Headings (MeSH) that describe the content of the articles. We use these to assign articles to particular fields. There are 27,149 raw terms in the 2014 MeSH vocabulary and they vary widely in their descriptive detail. For instance, some articles are tagged with general terms such as "Body Regions" and some are tagged with more detailed terms such as "Peritoneal Stomata". In order to construct comparable fields, we aggregate all

MeSH terms to a similar level of descriptive detail. This process—described in detail below and in Appendix C—leaves us with 6,159 aggregated MeSH terms.

To understand our aggregation method, first note that MeSH terms have a hierarchical structure. At the top of the hierarchy (1-digit terms) are 16 very general terms such as "Anatomy", "Organisms", and "Diseases". Beneath each of these 1-digit MeSH terms is a group of more detailed 2-digit MeSH terms. For instance, "Body Regions" is a 2-digit MeSH term beneath the 1-digit term "Anatomy". Beneath each 2-digit MeSH term is a group of even more detailed 3-digit MeSH terms. This structure continues to 12-digit MeSH terms. To reduce the amount of variation in the breadth of fields, we aggregate all MeSH terms to the 4-digit level. Aggregation is complicated by the fact that some more detailed (lower level) MeSH terms are associated with more than one higher-level 4-digit MeSH term. In these cases, we distribute (prorate) the weight of each higher-level MeSH term evenly across all of the 4-digit MeSH terms that are beneath it.

Once we have finished this aggregation process, we are able to transform each article's raw MeSH terms, which vary dramatically in terms of degree of aggregation, into 4-digit MeSH terms, which are considerably more uniform in terms of degree of aggregation. We then characterize the fields to which an article belongs by prorating the article equally across its 4-digit MeSH terms. Thus, each article is fractionally assigned to one or more 4-digit MeSH term fields. Appendix C, Figures C.1 and C.2 shows the distribution of the number of MeSH4 terms per article by publication year.

Field-Period Pairs

All metrics for high impact and transformative science (HITS) are defined for field-period pairs, i.e., a combination of a specific 5-year period and 4-digit MeSH term. The span of analysis, from 1983 through 2012, is divided into six consecutive 5-year periods, starting with 1983-1987 and ending with 2008-2012. Papers are sorted into these bins based on publication year. For each period, there exist 6,159 aggregated 4-digit MeSH terms. Six periods by 6,159 fields results in 36,954 period-field pairs, However, not all field-period pairs that have articles to generate values for all eleven metrics. Overall, 15,051 field-period pairs have articles and are used in this study. Table 2 shows the use (mentions) of highly used concepts (*BMentAll*) and the mean of forward citations (FCiteMean) for five relatively highly ranked fields. Gene Expression Profiling does not have any articles in two of the periods, but has the highest value for of *BMentAll* in the 2008-2012 period across the fields displayed. It is noteworthy that *BMentAll* increases over time because the number of concepts increases while FCiteMean declines in the latest years because the length of time over which citations can accrue is shorter, a factor we control below.

	1983-1987	1988-1992	1993-1997	1998-2002	2003-2007	2008-2012
Field						
BMentAll						
Gene Expression Profiling	1.00	.	.	4.98	6.22	6.73
Intracellular Signaling Peptides and Proteins	1.03	1.52	3.71	5.89	6.65	6.67
Neoplasm Proteins	1.12	1.75	3.06	4.41	5.38	5.46
RNA Viruses	1.13	1.99	2.69	3.33	3.67	3.73
Retroviridae Infections	1.36	2.60	3.12	4.06	4.61	4.85
FCiteMean						
Gene Expression Profiling	17.16	.	.	71.25	40.43	16.23
Intracellular Signaling Peptides and Proteins	43.22	65.07	92.85	67.18	40.01	15.89
Neoplasm Proteins	68.62	80.59	78.38	58.96	40.69	15.97
RNA Viruses	35.87	38.10	35.61	35.54	28.44	12.53
Retroviridae Infections	34.69	24.07	22.01	24.29	19.27	9.77

Table 2: Exemplary depiction of field-period pairs. All six time periods are shown but only five of the 6,159 fields.

4. Methods

Eleven metrics grouped by seven different dimensions of impact and transformativeness were introduced in Section 2 and they are defined and operationalized here.

For both the text-based and the citation-based metrics, we develop metrics to identify the impact and transformativeness of the articles published in a given field-period pair and refer to the articles, fields, and time periods for which we are measuring impact and transformativeness as the "target" articles, fields, and time periods. We refer to the articles, fields, and time periods over which we measure the impact and transformativeness of the target articles as the "measurement" articles, fields, and time periods. We note that depending on the metric, the measurement period may be before or after the target period—some of our metrics are forward looking while others are backward looking. Additionally, our measurement period and target period and field can overlap. Consider an article published in 1990. That article falls in the 1988-1992 target period. When we count forward citations to that article, we use citations occurring in articles published from 1990-2014, which includes three years of the 1988-1992 period. Similarly, when we generate backward citation ages for that article, we will get citations to articles in 1990 and all earlier years, which also includes three years of the 1988-1992 period.

To compute the text-based metrics, we begin with the full MEDLINE 2014 baseline files containing 22,376,811 articles published between 1809 and 2014. We index all words, word-pairs, and word-triplets that appear in the title or abstract of a MEDLINE article. We generically refer to these n-grams as "concepts". Next, we extensively process these concepts by eliminating stop words, stemming and lemmatizing each word, and applying a variety of other operations. See Appendix B for details.

After processing the MEDLINE corpus, we take the intersection of the MEDLINE and WOS database obtaining the 13,737,835 articles in Table 1. This set of articles contains 109,912,224 unique concepts. Next, we use article publication dates to identify the first year each concept is mentioned. We call this a concept's "vintage year". Further restricting our sample to the 10,778,696 articles published between 1983 and 2012, we obtain 95,393,331 concepts with vintage years between 1982 and 2012. Next, we count the number of times a concept is mentioned subsequent to its vintage year. To focus on the most important concepts, we identify the top .01 percent of concepts from each vintage (including all tied concepts in the case of ties at the threshold) — a total of 10,128 top concepts (including 589 due to ties) with vintages between 1983 and 2012. We use these top concepts to construct our text-based metrics.

Next, we verbally define each of the eleven metrics we have developed to capture high impact and transformative work. The full name of each metric and its variable name (in parentheses and italics) as well as formal definitions are given in Appendix D, Table D1. Summary statistics for all metrics, and all field-period pairs are presented in Table 3. This table also provides details on the number of field-period pairs for which each metric can be computed and info on which target periods and fields are associated with each of the metrics.

Radical–Generating

Top Concept Births (*Concepts*): To measure the generation of important new ideas, we measure how many of the top 10,128 concepts identified in the previous section are produced by a MeSH4 field in a particular period. To construct this metric, we first assign each concept to a period on the basis of its vintage. For instance, all concepts with a vintage between 2003 and 2007 are assigned to the 2003-2007 period. Second, we assign each concept to MeSH4 fields. To do this, we identify all articles that use a particular concept in the first year it was introduced (its vintage year) and then identify the MeSH4 fields of these articles. We then prorate the concept equally across these fields. Finally, we sum the number of top concepts assigned to each MeSH4 field-period pair. We call this sum the number of "top concept births". Concepts *are expected to be increasing with the radicalness of work.*

Top Concept Mentions (*BMentN*): To measure the utilization of important new concepts, we identify how many times one of the top 10,128 concepts identified in the previous section are used within N (N=0, 3, 5, 10, and all prior years) years of the concept's vintage. To construct this metric, we first identify all articles that use a top concept from any vintage. Second, we assign each article to a period on the basis of its publication year. For instance, all articles published between 1993 and 1997 are assigned to the 1993-1997--period pair. Third, we assign each article to MeSH4 fields by equally prorating the article across the fields with which the article is tagged. Fourth, we count the number of top terms introduced within the last *N* years used by each article. Finally, we sum across all articles assigned to each MeSH4 field-period pair. We call this sum the number of "top concept mentions". BMentN *are expected to be increasing with the radicalness of work.*

Radical–Destroying

Backward Citation Age (*BCiteAge*): This measure reflects the age of the works cited in articles. The age of a backward citation is the difference between the publication year of the citing article and the publication year of the cited article (backward citation). For each citing article, a mean backward citation age is constructed by averaging the ages of its backward citations. *BCiteAge* for a target MESH4 field and 5-year period is the average of the average backward citation age across all articles published in the target field -period pair. BCiteAge *decreases with destructive radicalness.*

Risky

Variance of Forward Citations (*FCiteVar*): In economics, the variance of the returns to an investment or asset are used as a proxy for the investment or asset's risk. Here the risky nature of articles published in a field-period pair is measured by the variance in forward citations they receive. This metric uses all subsequent years (including later years in the target vintage) without limitations. For example, suppose a target field-period contains three articles, each assigned exclusively to the target field. Now consider one case in which each article receives 30 citations, and a second case in which two articles receive no citations and one receives 90 citations. The field-period forward citation mean is the same in both cases (it is 30) but in the first case *FCiteVar* is 0 $\left(=\frac{1}{3}(3\times(30-30)^2)\right)$ and in the second case *FCiteVar* is 1800 $\left(=\frac{1}{3}(2\times(0-30)^2+(90-30)^2)\right)$. FCiteVar *ranges between 0 and infinity and increases with riskiness.*

Multidisciplinarity

Herfindahl of Backward Citations (*BHerfCite*): *BHerfCite* for a target MeSH4 field and 5-year period is an index of field dispersion of the articles cited by the articles published in that target field and period. *BHerfCite* is computed by squaring the total number of backward citations from each field, summing over the squares, and subtracting the result from 1. For example, if the articles published in a target field-period cited 1500 articles, 500 in each of three fields, the field-period's *BHerfCite* would be .667 $\left(=1-\left(\frac{1}{3}^2+\frac{1}{3}^2+\frac{1}{3}^2\right)\right)$. BHerfCite *ranges between 0 and 1 and increases with multidisciplinarity—the breadth of fields from which the article draws*.

Herfindahl Index of the Breadth of Existing Concepts Used (*BHerfMentN*): To measure the breadth of ideas that a field is drawing on, we use a Herfindahl index of the dispersion of the top concepts used by the articles published in that target field and vintage period. For a set of concepts from a given field and vintage period, we square each n-gram's share of total mentions. We then sum over the squares and subtract them from 1. We do this separately by the number of years (*N*=0, 3, 5, 10, and all prior years) since each concept was first used. BHerfMentN *ranges between 0 and 1 and increases with the breadth of ideas from which the field draws*.

Wide Impact

Herfindahl Index of Forward Citations (*FHerfCite*): *FHerfCite* for a target MESH4 field and 5-year period is an index of field dispersion of the articles citing the articles published in that target field and vintage. *FHerfCite* is computed by squaring the share of forward citations from each field, summing over the squares, and subtracting the result from 1. For example, if the articles published in a target field-period were cited by 1500 articles, 500 in each of three fields, the field-period's *FHerfCite* would be .667 $\left(=1-\left(\frac{1}{3}^2+\frac{1}{3}^2+\frac{1}{3}^2\right)\right)$. FHerfCite *ranges between 0 and 1 and increases with breadth of impact*.

Herfindahl Index of the Breadth of the Future Use of Concepts Introduced (*FHerfMent*): To measure the breadth of use of the concepts generated in a field, we use a Herfindahl index of the dispersion across fields in the future use of the concepts introduced in a field-period pair. For a set of concepts from a given field and vintage period, we take the share of mentions in subsequent periods across all fields, square each field's share of total mentions. We then sum over the squares and subtract them from 1. FHerfMent *ranges between 0 and 1 and increases with breadth of use of the concepts generated in a field.*

Growing Impact

Forward Citation Age (*FCiteAge*): This measure captures the typical length of time between when works are published and citations to that work occur. The age of a forward citation to a cited article is the difference between the publication year of the citing article (forward citation) and the publication year of the cited article. For each cited article, a mean forward citation age is constructed by averaging the ages of its forward citations. *FCiteAge* for a target MeSH4 field and 5-year period is the average of the article averages of forward citation age across all cited articles published in the target field-period. FCiteAge *increases with growth of impact.*

High Impact

Mean Forward Citation Count (*FCiteMean*): *FCiteMean* for a target MESH4 field and 5-year period is the average forward citation counts across all articles (including those that receive no citations) published in the target field and period. FCiteMean *increases with impact.*

Forward Citation Percentile (*FCite*N): This series of metrics captures the impact as measured by forward citation counts at various percentiles of the distribution of forward citations. Formally, we rank articles in a target field-period pair by their forward citation counts (including those that receive no citations). *FCite*N is the forward citation count below which *N* percent of articles in a target field-period pair are found. For example, *FCite75* is the forward citation count of the article at the 75th percentile for the target field and period. *FCiteN* is constructed for N = 25, 50, 75, 90, 95, 99, 99.9 and 99.99. FCite*N* *captures the impact of the most cited articles in a target field-period and increases with impact.*

As indicated, articles may be assigned to more than one MeSH category. In calculating each metric for each MeSH category, we weight articles by the share of the article falling in that MeSH category.

Table 3: Summary Statistics for all Metrics for time periods 1983-1987, …, 2008-2012 and all MESH4 Fields in MEDLINE.

Variable Name	Mean	S.D.	Measurement Term and Fields
Citation-Based Metrics			
FCiteMean	22.309	12.670	All subsequent years (including later years in the target vintage) through 2014. All of MEDLINE.
FCite25	3.489	2.656	All subsequent years (including later years in the target vintage) without limitations. No limitations on measurement fields. Includes both MEDLINE and non-MEDLINE indexed articles.
FCite50	9.838	6.137	
FCite75	23.767	13.366	
FCite90	50.264	27.736	
FCite95	78.926	44.069	
FCite99	192.525	111.922	
FCite99.9	586.390	393.483	
FCite99.99	1626.023	1623.863	
FHerfCite	0.979	0.005	All subsequent years (including later years in the target vintage) through 2014. All of MEDLINE.
FCiteAge	5.229	2.555	All subsequent years (including later years in the target vintage) through 2014. All of MEDLINE.
BHerfCite	0.979	0.006	All previous years (including earlier years in the target vintage) without limitations. All of MEDLINE.

BCiteAge	9.642	2.605	All previous years (including earlier years in the target vintage) without limitations. No limitations on measurement fields. Includes both MEDLINE and non-MEDLINE indexed articles.
FCiteVar	4803.615	14596.430	All subsequent years (including later years in the target vintage) without limitations. No limitations on measurement fields. Includes both MEDLINE and non-MEDLINE indexed articles.
Text-Based Metrics			
Concepts	32.235	72.099	1983-1987,…, 2008-2012; All MESH4 Fields in MEDLINE.
BMent0 BMent3 BMent5 BMent10 BMentAll	0.003 0.037 0.085 0.294 1.028	0.006 0.033 0.078 0.258 0.930	1983-1987,…, 2008-2012; All MESH4 Fields in MEDLINE.
FHerfMentions	0.996	0.008	1983-1987,…, 2008-2012; All MESH4 Fields in MEDLINE.
BHerfMent0 BHerfMent3 BHerfMent5 BHerfMent10 BHerfMentAll	0.910 0.969 0.976 0.981 0.983	0.184 0.066 0.053 0.047 0.046	All articles that use a concept within N (N=0, 1, 3, 5, all years) years of the concept's vintage and published during the 1983-1987,…, 2008-2012 periods. All MESH4 Fields in MEDLINE.

Note. 15,051 Field-Period pairs.

5. Results: Comparison of Metrics

The eleven general metrics were analyzed and compared using a factor analysis to identify different dimensions of transformativeness. Because the forward citation rates are the conventional measure of impact or influence, we perform a factor analysis on impact metrics as a group.

When conducting the factor analysis, we first compute the natural logarithm of one plus all metrics and then take deviations from field and period means. We do this by regressing the natural logarithm of one plus each variable on a set of dummy variables for 4-digit MESH field and a set of dummy variables for 5-year period. Formally, let M_{fp} denote the value of a metric for field *f* in period *p*; $\overrightarrow{F_{fp}}$ denote a vector of field dummy variables (or fixed effects) equal to one for field *f* and zero for the other fields; and $\overrightarrow{P_{fp}}'$ denote a vector of period dummy variables (fixed effects) equal to one for period *p* and zero otherwise. We estimate,

$$ln(M_{fp}+1) = \overrightarrow{F_{fp}}'\beta + \overrightarrow{P_{fp}}'\gamma + \epsilon_{fp} \,.$$

We analyze the residuals from this equation, ϵ_{fp}. Taking the natural logarithm of the metrics reduces weight on variations in the right tail of the metrics, which tend to be highly right-skewed (adding 1 is a commonly used approach to address values of zeros). The period dummy variables address the end of our outcomes data in 2014, with the various vintages being different lengths of time from being truncated. Eliminating such cross-period variation when estimating factor loadings means that changes in the overall level of the metrics from period to period do not influence the factor loadings. We run the same comparison dropping the 2008-2012 period and found similar results. The field dummy variables address differences across fields in characteristics such as size. Thus, large fields are likely to generate more citations and the concepts they originate are likely to be more heavily mentioned. Eliminating cross-field variation in estimating the factor loadings means that the metrics are not influenced by overall differences across fields.

For all subsequent analyses, the dataset comprises all 15,051 field-period pairs and observations are weighted by the number of articles in that field-period pair.

Figure 2 reports results from the factor analysis for the three composite metrics – Radical Generative work, which combines *Concepts* and *BMentN*; Multidisciplinarity, which combines *BHerfCite* and *BHerfMentN*; and *Impact*, which combines *FCiteMean* and *FCiteN*. Note that the metrics for Radical Destructive work (*FCiteAge)*, Risky work (*FCiteVar*), and Increasing Impact (*FCiteAge*) are all generated from a single metric, so that no factor analysis was performed. The metric for Breadth of Impact is based on only 2 metrics (*FHerfCite* and *FHerfMent*) and the factor analysis is not plotted. In all cases, the first factor accounts for the vast majority of the variation (79%-94%) and is the focal point here.

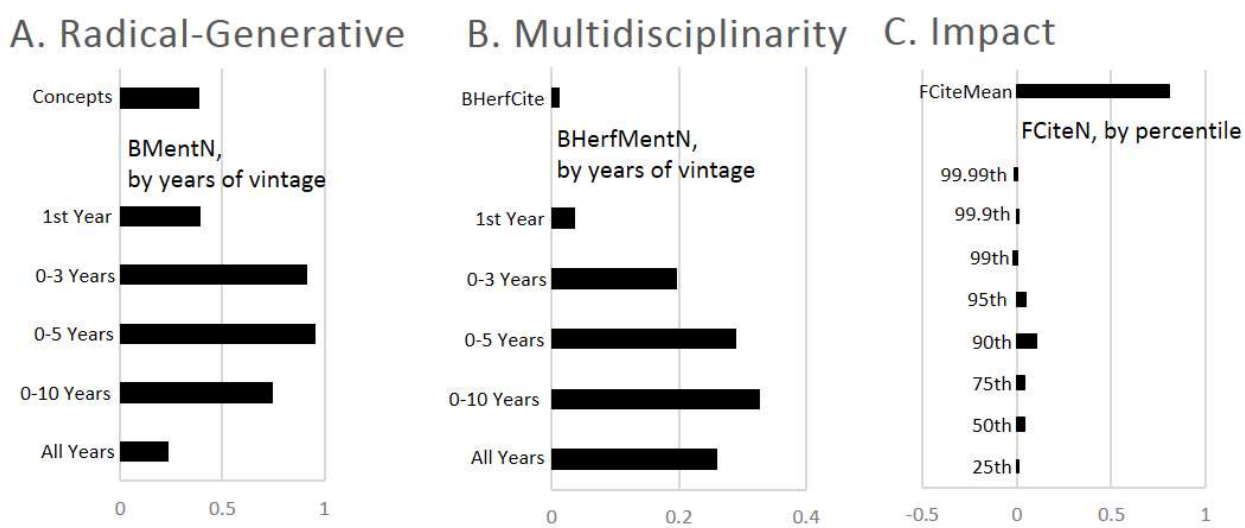

Figure 2. Factor analysis results for three of the seven metrics types. Share of variation explained by first factor is .790 for Radical-Generative work.943 for Multidisciplinarity; and .853 for Impact.

Radical-Generative science (Panel A) loads positively on the birth of new concepts (*Concepts*) and the mentions of important concepts of various ages (*BMentN*), with the highest loading on concepts that are 0-5 years old. Thus, concepts that are older receive less weight than those that are 0-5 years old, while those that are younger receive more weight because they appear in multiple groups.

Multidisciplinary science (Panel B) loads positively on the dispersion of citations (*BHerfCite*) and the dispersion in the use of top concepts (*BHerfMentN*), both measured using Herfindahl indices. The loading on concepts that are 0-10 years old is the highest, but concepts in their first year since origin is the highest because these concepts are included in the other age categories, so the dispersion of the use of the newest concepts is particularly related to multidisciplinarity.

Our metric for high Impact science (Panel C) is generated from the mean of forward citations (*FCiteMean*) and quantiles of the distribution of forward citations *FCiteN* (here constructed for N = 25, 50, 75, 90, 95, 99, 99.9 and 99.99), which tend to be closely related. The first factor of impact accounts for 85% of the variation. Mean Citations (*FCIteMean*) has the highest factor loading. Interestingly, the factor loadings on the quantiles of the forward citation distribution increase from the 25th percentile of the citation distribution through the 90th percentile and then decline, so that the lowest factor loadings for the 99th and the 99.99th percentile of the citation distribution are slightly negative (the 99.9th is positive, but small, .00663). We show below that these papers that are most likely to be transformative.

Table 4 reports correlations between the various aspects of impact and transformativeness. The results show that many aspects of transformativeness are positively correlated, indicating some real cohesion of these metrics of transformativeness. The Radical-Generative and Risky metrics are comparatively highly correlated (*ρ=.271*), suggesting that they are capturing inter-related phenomena. Both metrics are strongly positively correlated with High Impact (*ρ=.383* and *.556*, respectively). Multidisciplinarity and Wide Impact research are also comparatively highly correlated (*ρ=.246*). It is intuitive and reassuring that work that draws on a wide range of work itself draws on by a wide range of work.

Other aspects of transformativeness appear to be only weakly related or unrelated. Radical-Generative is essentially uncorrelated with Wide Impact across disciplines (*ρ*=.067). Radical-Destructive is only weakly correlated with Radical-Generative (*ρ=.105*) suggesting that the generation of important new concepts in a field frequently occurs without rendering old science obsolete. Radical-Destructive is essentially uncorrelated with Risky (*ρ=.050*) and Wide Impact (*ρ=-.017*), but weakly negatively correlated with Multidisciplinarity (*ρ=-.096*). Interestingly, Multidisciplinarity is also not strongly correlated with High Impact (*ρ =.100*) or Radical-Generative (*ρ =-.076*). These correlations contrast with the perspective that work that brings together differing scientific approaches or viewpoints generates more influential and radical scientific output.

Table 4. Correlations between individual metrics of Impact (labeled "Impact") and Transformativeness (the other six metrics).

	Radical - Generative	Radical - Destructive	Risky	Multidis-ciplinary	Wide Impact	Growing Impact	High Impact
Radical - Generative	1						
Radical - Destructive	0.1047	1					
Riskiness	0.2714	0.05	1				
Multi-Disciplinary	-0.0756	-0.0961	0.0764	1			
Wide Impact	0.0672	-0.0167	0.0841	0.2467	1		
Growing Impact	-0.2941	-0.3528	-0.2318	-0.0348	-0.2424	1	
Impact	0.3832	0.0271	0.5561	0.0999	0.0342	-0.1995	1

The strongest correlation observed in Table 4 is between High Impact and Risky (*ρ=.556*). We see a strong correlation between the variance in forward citation counts and citations at *all* quantiles of the citation distribution, including the 25th percentile and the median, for example. This suggests the possibility of a trade-off between risk and return in scientific research.

As indicated, growth of impact over time, measured by the average time to citations, is negatively correlated with all the other metrics, which contrasts with views that transformative work takes a long time to have an impact. We have broken forward citations to the work in each field-period pair into those arising in the first five years since publication and those arising six or more years since publication. Both metrics are positively related to each aspect of impact and transformativeness, but citations in the first 5 years are more strongly correlated with the other metrics for transformativeness (and impact) than are citations six or more years out - the correlation between transformativeness and citations in the first five years is .674 while that between transformativeness and citations six or more years after publication is only .247. Put differently, transformative work is heavily cited in the long run, but it is even more heavily cited in the short run. A limitation of the study is that we cannot measure impact over very long time periods, thus we cannot rule out the possibility that the most transformative work grows in impact over much longer time horizons, e.g. over many decades.

Figure 3 shows the first factor from a factor analysis of the six metrics of transformativeness (which accounts for 107% of the variation). High Impact was excluded to assess how impact and transformativeness are related. Transformativeness loads positively on all of the metrics but Growing Impact, suggesting that transformativeness represents a cohesive construct.

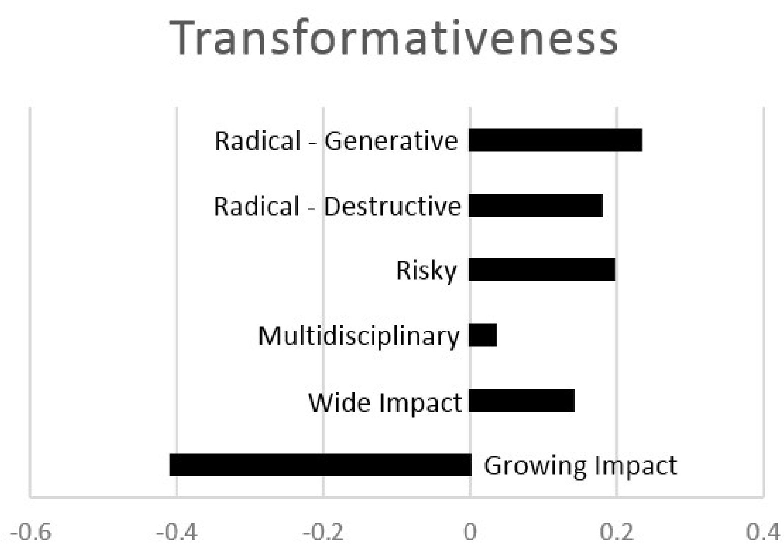

Figure 3. Factor analysis of six different HITS metrics, High Impact is omitted.

Figure 4 relates the seven HITS metrics to impact and transformativeness. Aside from Growing Impact, all the metrics of transformativeness are positively related to both impact and transformativeness. The metric for High Impact has a correlation with itself of 1 (by construction) and a correlation with transformativeness of .401. Thus, while impact and transformativeness are positively related, they also seem to constitute distinct phenomena. The Risky metric is most strongly correlated with Impact. Radical-Destructive and Wide Impact are both strongly related to transformativeness, but essentially unrelated to impact. Lastly, Growing Impact is strongly negatively related to transformativeness and somewhat negatively related to Impact. Interestingly, the correlation between impact and citations in the first five years is .881, falling slightly to .762 for citations six or more years after publication.

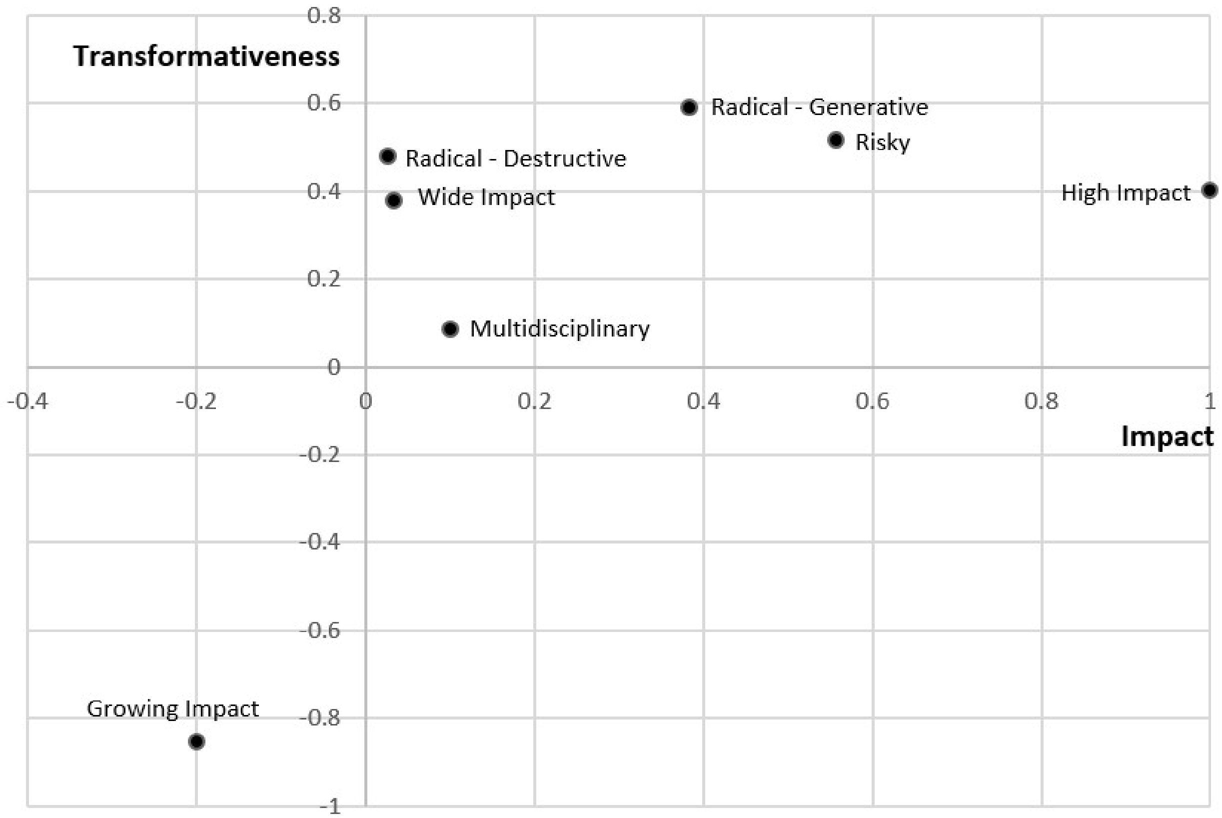

Figure 4. All seven HITS metrics related to Impact and Transformativeness

Figure 5 shows how the various forward citation metrics (*FCiteMean* and *FCiteN*) relate to impact and transformativeness. As indicated in Figure 2C, impact loads most heavily on mean citations and citations at the 90th percentile with weight declining (or going negative) at the highest and lowest percentiles of the citation distribution. The correlations between the impact measure and the highest percentiles of the citation distribution are positive even though the factor loadings are negative because all of the citation metrics are positively correlated. It is intuitive that transformative works should be exceptionally highly cited. Indeed, the strength of the relationship between the percentiles of the citation distribution and transformativeness increase monotonically up to the 99th percentile of the citation distribution (compared to the 90th percentile for citations) and then decline moderately to the 99.99th percentile. Strikingly the 99.99th quantile of the citation distribution is almost as strongly related to transformativeness as it is to impact. These results suggest that the most cited impactful works reflect a distinct phenomenon from other high impact work and that they are the most likely to be transformative.

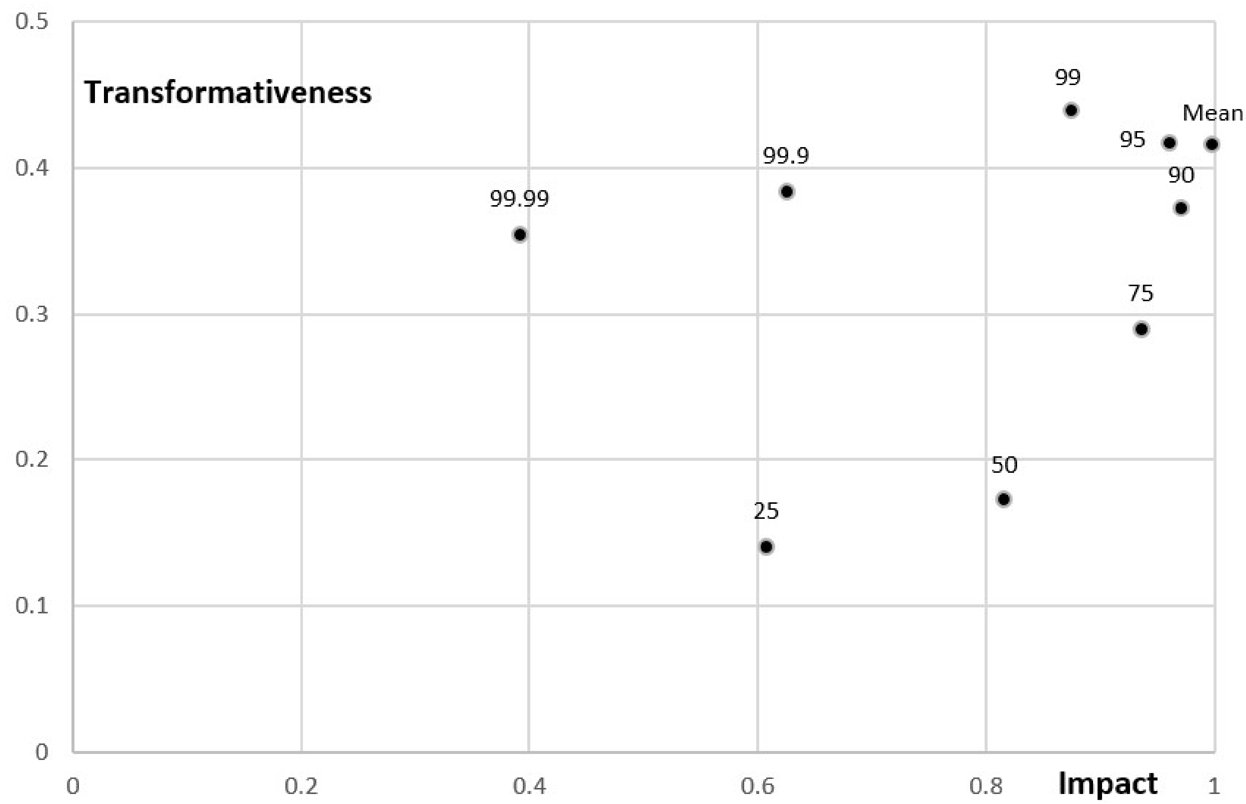

Figure 5. FCiteN related to Impact and Transformativeness

To provide some summary of our analysis, we take the factor loadings from our factor analysis and use them to generate the impact and transformativeness metrics for each 4-digit MESH fields, see Figure 6. In doing so, we average across all of the periods from 1982-2012. While there are differences across fields, we do not eliminate field differences for this analysis. The figure shows a strong positive relationship between impact and transformativeness, but also differences. For instance, gene and expression profiling, retroviridae infections, and RNA viruses all score considerably higher on transformativeness relative to impact. Interestingly, there are many low impact fields and fields that are also rather low in transformativeness. Many relate to historic or societal factors rather than to biomedical conditions, treatments or technologies.

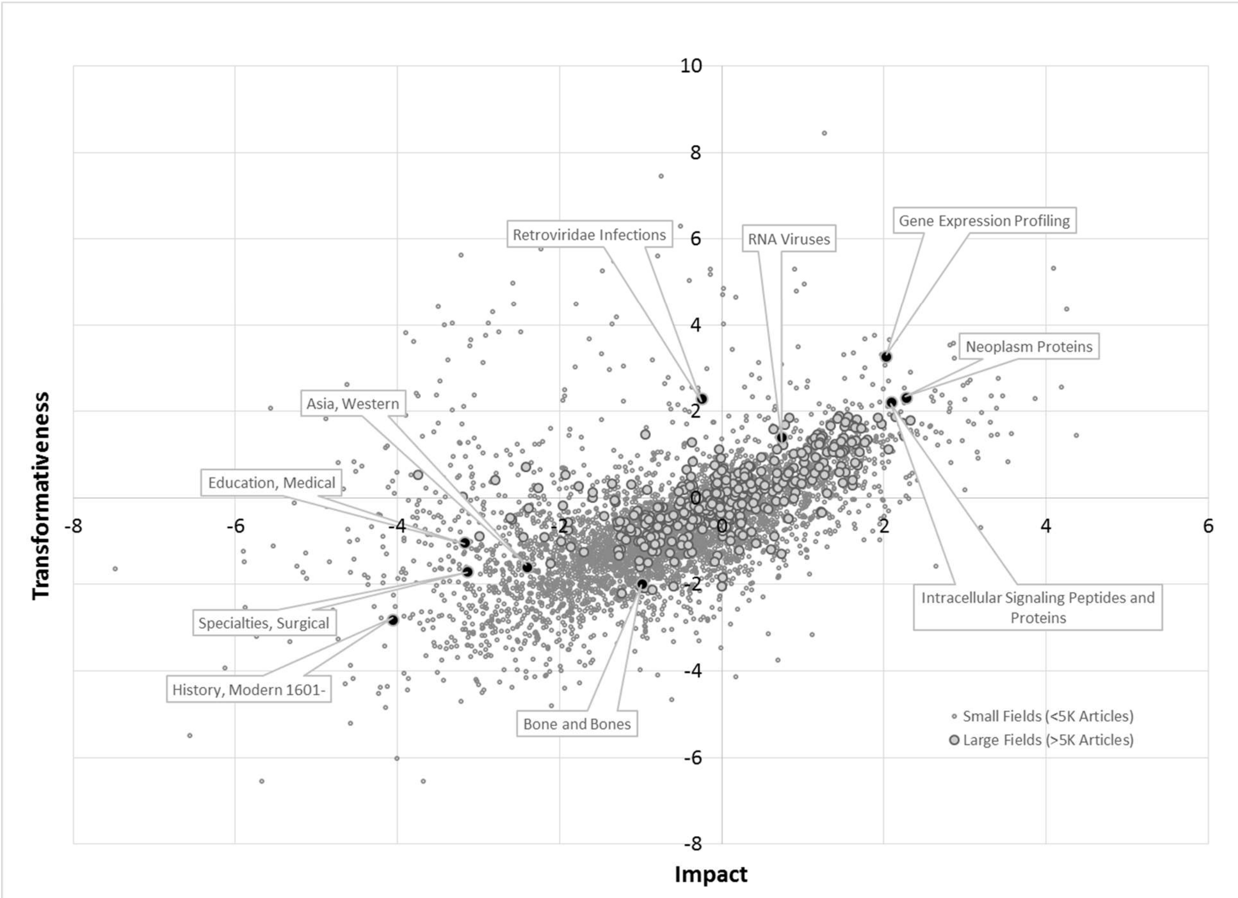

Figure 6. Ranking of fields in terms of impact and transformativeness across all periods (1982-2012). Field size determined by the number of (weighted) articles across all periods.

Just as we preserve the cross-field variation in estimating the field rankings, we have rerun the factor analysis preserving the cross-field variation. The results are broadly similar to those reported above, with all the components of transformativeness entering in the same way as above. One clear difference is that the correlation between transformativeness and impact is higher when cross-field variation is preserved (*ρ=.668* versus *.401*, when the cross-field variation is eliminated). This result is intuitive, in that it indicates that differences in transformativeness across fields are more strongly related to impact than are changes in transformativeness within fields over time. Put somewhat differently, the fields that are transformative tend to be more impactful while fields that are temporarily more transformative experience smaller increases in impact.

6. Discussion

The science policy community is increasingly focusing on transformative research, yet there are few metrics to identify it and, ironically, the related concept of revolutionary science is falling out of favor. Drawing on existing conceptualizations of transformative research, this paper presented eleven metrics of transformative research. Specifically, transformative research is viewed as being radical, both generating important new ideas and destroying existing ideas; multidisciplinary and impacting a wide range of disciplines; risky; having a wide and growing impact over time; and having a high impact.

The metrics were exemplarily applied to 15,051 fields of biomedical research over six five-year periods from 1983-1987 to 1988-2012. Our validation studies show that the metrics behave as expected. Transformativeness and impact are positively correlated (*ρ=.401*), but distinct. The interrelations between specific metrics of transformativeness are often positively related but some are only weakly related or unrelated. Metrics of the use of wide-ranging ideas or multidisciplinarity are closely related to the breadth of impact. A notable exception we find is that the growth of citations is negatively related to transformativeness—while citations six or more years after publication are increasing in transformativeness, citations within the first five years increase even more. Whether this represents the limitations of the timespan of our data or a fundamental fact of transformative research, we leave to future research. In addition, we find that the displacement of old science coincides with the generation of radical new science only moderately, and that neither correlates strongly with multidisciplinarity. Interestingly we find a strong positive association between impact and riskiness which suggests the possibility a trade-off between risk and return in scientific research.

Decision makers have a number of choices when selecting metrics, with individual choices depending on data access, preferences, but also expertise and computational resources. All metrics for MEDLINE articles introduced in this paper are freely available for scholarly research subject to licensing restrictions (in the case of proprietary citation data). For those interesting in generating our metrics over their own data or corpa, the titles and abstracts necessary to generate text-based metrics are openly available. The metrics of new concept births and mentions of concepts are relatively easy to compute, making it possible for anyone to compute metrics of radical generative work. The backward and forward Herfindahl indices of the breadth of mentions of new concepts have the same data requirements but are computationally more demanding. Thus, our text-based metrics of radical generative research, breadth of impact, and multidisciplinarity should be accessible to most practitioners. Generating citation metrics requires a different type of data access, e.g., to Thomson Reuters' Science Citation Index as discussed here, or to one of the other citation databases. Calculating the mean citations to the works in a field-period pair, the quantiles of the citation distribution, and the variance of citations across the works in a field-period pair require total citation counts to articles exclusively and are not computationally demanding. These provide good measures of impact and riskiness. As indicated, the extreme right tail of citations (e.g., the 99.99^{th} percentile of the citation distribution) is relatively strongly related to transformativeness and is not computationally burdensome either. The other citation metrics require data that go beyond raw forward citation counts, namely data on citing-cited article pairs. Backward citation ages and forward citation ages are both straightforward computationally, providing metrics of radical destructiveness and growth of impact. As with the text variables, the forward and backward citation Herfindahl's are more computationally burdensome. Thus, while users must generate the metrics that suit their data access and computational environment, the choices they face when implementing our methods are obvious.

There are a number of limitations related to the data used and the metrics defined in this study. First, all of our analyses are limited in topical focus to articles published in MEDLINE and are limited temporally to the period 1983-2012. In terms of citation data from the Web of Science corpus, we used all backward citations of articles published in 1983-2012 yet did not have access to forward citations beyond May 20, 2014. We used MEDLINE's titles and abstracts for the text-based metrics exclusively.

We used the MeSH hierarchy to measure the breadth of knowledge used in our target articles and the breadth of utilization of the ideas generated by our target articles and hence our definition of fields and our text-based metrics are restricted to the MeSH classification of the MEDLINE corpus. MEDLINE mostly covers biomedical research, a study of other research disciplines with different publication norms, researcher team sizes, and funding opportunities might provide different results. Given that our last period ends in 2012 and citation data ends in 2014 it is highly probable that the impact of some articles is still materializing. However, the results of our analysis seem robust as omitting the 2008-2012 period does not change values dramatically.

There are a number of directions for future work. First, we look to validate our metrics in a variety of ways. Specifically, we are implementing an online interactive interface that users can visit to identify highly transformative research by selecting any of the eleven metrics and associated parameter values. This will make it possible to solicit feedback on the accuracy, representativeness, and usefulness of the metrics from subject matter experts. We are also working to identify and engage subject matter experts in specific research to validate our metrics individually.

Beyond validation, there are a number of other important avenues for future research. First, we are interested in identifying the factors—from the funding mechanisms, to the demographics of the researchers in fields, to the networks of researchers—that lead to the production of transformative science. Second, we seek to attach analogous metrics to individual articles, not just to entire research fields in a given period as we have done here. Such estimates would allow us to identify specific transformative works. They would also allow us to identify features of research teams that are likely to produce transformative work.

Acknowledgements

We thank Mikko Packalen for helpful input on the text analysis and Neil Smalheiser and Vetle Torvik for the use of their Author-ity data and Thomson Reuters for access to Web of Science records. Sara Bouchard provided design expertise for optimizing the figures. All authors gratefully acknowledge support from P01 AG039347. Weinberg thanks NSF 1064220, 1348691, 1535399 and the Ewing Marion Kauffman and Alfred P. Sloan Foundations. Weinberg was supported on P01 AG039347 by the NBER directly and on a subaward from NBER to Ohio State. Börner and Light are partially supported by the National Science Foundation under award EAGER 1566393 and NCN CP Supplement 1553044 plus the National Institutes of Health under award U01CA198934. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Acemoglu, Daron; Ufuk Akcigit, and Murat Celik. 2015. "Young, Restless, and Creative: Openness to Disruption and Creative Innovations." NBER Working Paper No. 19894.

Bettencourt, Luis M. A.; David I. Kaiser; and Jasleen Kaur. 2009. "Scientific Discovery and Topological Transitions in Collaboration Networks." *Journal of Informetrics*.

Bollen J, Van de Sompel H, Hagberg A, Chute R, 2009 A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE* 4(6): e6022. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0006022

Chen, C. Julian. 2007. *Introduction to Scanning Tunneling Microscopy*. Oxford: Oxford University Press.

Cronin, B. & Sugimoto, C.R. (Eds.) (2015). *Scholarly Metrics Under the Microscope: From Citation Analysis to Academic Auditing*. Medford, NJ: Information Today, Inc./ASIST.

Cronin, B. & Sugimoto, C.R. (Eds.) (2014). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, MA: MIT Press.

Evans, James and Jacob G. Foster. 2016. "Measuring Novelty by Simulating Discovery." Working Paper.

Funk, Russel J. and Jason Owen-Smith. Forthcoming. "A Dynamic Network Measure of Technological Change." *Management Science.*

Gau, Xue and Zhang, Pumin. 2007. "Transgenic RNA Interference in Mice". Physiology 22: 161-166.

Hanning, Guo, Scott B. Weingart, and Katy Börner. 2011. "Mixed-indicators model for identifying emerging research areas". *Scientometrics* 89 (1): 421-435.

Hutchins, B. Ian; Xin Yuan; James M. Anderson; and George M. Santangelo. "Relative Citation Ratio (RCR): A New Metric that Uses Citation Rates to Measure Influences ta the Article Level." bioRxiv, doi: http://dx.doi.org/10.1101/029629.

Jones, Benjamin F. and Bruce A. Weinberg. 2011. "Age Dynamics in Scientific Creativity." *Proceedings of the National Academy of the Sciences* 108 (November 22): 18,910-14.

Thomas S. Kuhn (1947) *The Structure of Scientific Revolutions*, University of Chicago Press.

National Institutes of Health, 2015. *Transformative Research Award Program.* https://commonfund.nih.gov/tra/description. Accessed: 2016/08/01.

National Library of Medicine, 2016. *MEDLINE®PubMed® XML Element Descriptions and their Attributes.* https://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html. Accessed 2016/08/01.

National Science Board (NSB). 2007. *Enhancing Support of Transformative Research at the National Science Foundation* (NSB-07-32), National Science Foundation, Arlington, VA 22230.

National Science Foundation, 2015. *Definition of Transformative Research"* http://www.nsf.gov/about/transformative_research/definition.jsp, Accessed: 2015/09/03.

Stokes, Donald E. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation.* Washington, DC: Brookings Institution Press.

Wang, Dashun, Chaoming Song, and Albert-Laslo Barabasi. 2013. "Quantifying Long-Term Scientific Impact." *Science* (Vol 342, October 4): 127-132.

Wang, Jian; Reinhilde Veugelers; and Paula Stephan. 2016. "Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators." NBER Working Paper NO. 22180.

Appendix

A. Types and Metrics of Transformative Research

The below table identifies different types of transformative research and maps them to eleven metrics

Table A1. Types and metrics of transformative research

National Science Board	NIH Roadmap	NSF	NIH Common Fund
1. Generates New Paradigms and Scientific Fields			
New paradigms or new scientific fields generating new ones [paradigms]	Create … fundamental paradigms	Leads to the creation of a new paradigm … provides pathways to new frontiers.	Create … fundamental paradigms, bold, paradigm-shifting
Metrics of generating important new concepts: Concepts (*Concepts*)—Introduction of heavily-used new concepts Mentions (*BMentN*)—Use of important new concepts			
2. Radical, Destroying Existing Paradigms			
Radically different approaches or interpretations; overthrowing entrenched paradigms	Overturn fundamental paradigms	Radically change our understanding of an important existing scientific or engineering concept	Overturn fundamental paradigms
Metrics of obsolescence of existing knowledge: Backward Citation Age (*BCiteAge*) (negatively related)—The age of prior knowledge drawn on			
3. Risky			
	Inherently risky		Inherently risky
Metrics of variation of new utilization of the knowledge generated: Variance of Forward Citations (*FCiteVar*)—Variation in utilization of knowledge produced			
4. Multidisciplinary			
Often can be multidisciplinary			
Metrics of drawing on a wide range of knowledge Backward Herfindahl of Citations (*BHerfCite*)—Breadth of articles used Backward Herfindahl of Concepts (*BHerfMentN*)—Breadth of important concepts used			
5. Wide Impact			
	Impact a broad area of biomedical research		
Metrics of knowledge being used widely: Forward Herfindahl Citations (*FHerfCite*)—Herfindahl index of breadth of forward citations Forward Herfindahl Mentions (*FHerfMent*)—Herfindahl index of breadth of future use of concepts introduced			
6. Growing Impact			
Often fragile in their early stages			
Metrics of time path of utilization of knowledge: Forward Citation Age (*FCiteAge*)—Age of citations			
7. High Impact			
Scientific understanding	Profoundly impact …		

advances dramatically	biomedical research		
Metrics of citation impact: Forward citation counts (*FCiteMean, FCiteN*)—Number of citations			

B. Extracting and Processing Text

We first extract the title and abstract from each of the 22,376,811 records (articles) indexed in the 746 MEDLINE 2014 Baseline XML Files including article identifying information such as PMID and Version. After extracting the title and abstract, we then index all words, word pairs and word triplets (1-, 2-, and 3-grams). We then process the n-grams contained in these titles and abstracts by performing the following operations:

1. Convert all text to lower-case.
2. Eliminate 2- and 3-grams with words that cross the following characters: ,.?!;:)(\}\{][--.
3. Eliminate all remaining characters that are not alphanumeric.
4. Eliminate all n-grams that contain words appearing in the stopword list provided by the NLM at this address:
5. Eliminate all n-grams that contain the following character sequences: web, www, http, pubmed, MEDLINE.
6. Eliminate all n-grams that contain more than two adjacent numbers.
7. Eliminate all n-grams that have a length of less than three characters.
8. Keep all 1-grams with character length 3-29, 2-grams with character length 7-59, and 3-grams with character length 11-89.

   Next, we stems each word from each n-gram using the module *Lingua::Stem* from the Comprehensive Perl Archive Network (CPAN). Finally, we index all the processed n-grams from each title and abstract into 746 tab-delimited text files corresponding to the 746 MEDLINE XML files.

As discussed in Section 3, only the 15,916,023 articles published in 1983-2012 are used to compute the text-based metrics. Here we focus on the 10,778,696 articles available in Thomson Reuters' Science Citation Index Expanded™ that match to MEDLINE.

C. Aggregating MeSH Terms to Construct Fields

We use the Medical Subject Headings (MeSH) that tag most articles in MEDLINE to characterize the fields to which each article belongs. There are 27,149 raw terms in the 2014 MeSH vocabulary and they vary widely in their descriptive detail. For instance, some articles are tagged with general terms such as "Body Regions" and some are tagged with more detailed terms such as "Peritoneal Stomata". Thus, in order to construct comparable fields, we aggregate all MeSH terms to a similar level of descriptive detail.

To understand our aggregation method, first note that MeSH terms have a hierarchical structure. At the top of the hierarchy (first-level terms) are 16 very general terms such as "Anatomy", "Organisms", and "Diseases" Each of these 16 first-level terms are identified by a unique capital letter. For instance, "Anatomy" is identified by the letter "A", "Organisms" is identified by "B", and so on. Beneath each of these first-level MeSH terms is a group of second-level MeSH terms. For instance, "Body Regions" is a second-level MeSH term beneath the top-level term "Anatomy". Each second-level MeSH term is identified by the capital letter of the first-level MeSH term it is beneath and by two numbers. For instance, “Body Regions” is identified by “A01”. Beneath each second-level MeSH term is a group of third-level MeSH terms identified by the capital letter of the first-level term it is beneath, the two numbers of the second-level term it is beneath, and three subsequent numbers. For instance, “Anatomic Landmarks” is a third-level MeSH term under “Body Regions” and is identified as “A01.111”. This structure continues to depths of up to 12 levels.

Aggregating MeSH terms (that is, classifying lower level MeSH terms as a part of higher level MeSH terms) is not straightforward because some MeSH terms are beneath more than one higher level MeSH term and some articles can be tagged with multiple MeSH terms. Appendix Figure 2 illustrates our approach. Our handling of multiple levels is best illustrated by example. Consider the MeSH term “Dementia”. This MeSH term has two separate identifiers: “C10.228.140.380” and “F03.087.400”. Thus, this MeSH term falls under the first-level MeSH terms “Diseases” (identified by “C”) and “Psychiatry and Psychology” (identified by “F”). It also falls under the second-level terms “Nervous System Diseases” (“C10”) and “Mental Disorders” (“F03”). The problem arises because MEDLINE records only contain the MeSH terms themselves, not their identifiers. For instance, if a MEDLINE record is tagged with the MeSH term “Dementia”, we would not know if it was the “Dementia” that is beneath “Nervous System Diseases” or “Mental Disorders”.

If we wanted to aggregate all MeSH terms to the second-level, we would have to find a way to split “Dementia” between “Nervous System Diseases” and “Mental Disorders”. We opt for the straightforward method of assigning half to each higher level term. If we wanted to aggregate all MeSH terms to the fourth-level, “Dementia” would fall under the fourth-level term “Brain Diseases” (identified by “C10.228.140”) and “Dementia” itself (identified by “F03.087.400”). In this case, we still assign half of “Dementia” to “Brain Diseases” and half to “Dementia” itself. Finally, if we wanted to aggregate all MeSH terms to the fifth-level, “Dementia”, which has two fourth-level identifiers (“C10.228.140” and “F03.087.400”), is at a higher level of aggregation than the fifth-level. We deal with this by simply eliminating “Dementia”. It is too highly aggregated for our purposes.

Our last step is to apportion each article indexed in MEDLINE to the newly aggregated MeSH terms. Again, examples are the most illustrative. Suppose we come across an article that is tagged by “Dementia” and we want to aggregate to the fourth-level. We know from above that half of Dementia would be assigned to “Brain Diseases” and half would be assigned to “Dementia” itself. However, suppose (as is usually the case) that this article is also tagged with other MeSH terms. Specifically, suppose that the article is also tagged with the fourth-, fifth- and ninth-level MeSH term “Health Information Exchange” (identified by “L01.700.253”, “L01.399.500.500”, “L01.313.500.500”, and “E05.318.308.940.968.625.500.500”). By the process discussed above, one fourth of “Health Information Exchange” will be assigned to each of the four fourth-level MeSH terms: “Health Information Exchange” itself (“L01.700.253”), “Health Information Management” (“L01.399.500”), “Medical Informatics” (“L01.313.500”), and “Data Collection” (“E05.318.308”).

We assume that each of the original MeSH terms, “Dementia” and “Health Information Exchange” receive equal weight in characterizing the article. Under this assumption, the article will be apportioned to each fourth level MeSH term as follows:

1. 1/2*1/2=1/4 to “Dementia”
2. 1/2*1/2=1/4 to “Brain Diseases”
3. 1/2*1/4=1/8 to “Health Information Exchange”
4. 1/2*1/4=1/8 to “Health Information Management”

5. 1/2*1/4=1/8 to “Medical Informatics”
6. 1/2*1/4=1/8 to “Data Collection”

Obviously 1/4+1/4+1/8+1/8+1/8+1/8=1. Thus, the article that was originally tagged by the two MeSH terms “Dementia” and “Health Information Exchange” is now apportioned between six different fourth-level MeSH terms. In general, we apportion each MEDLINE article across aggregated MeSH terms in two stages. First, we equally apportion the original MeSH terms across the higher-level MeSH terms of which they are a part (e.g. apportion “Dementia” equally across “Brain Diseases” and “Dementia”). Second, we weight this apportionment by the inverse of the number of original MeSH terms that tag the article (e.g. our hypothetical article was tagged by two original MeSH terms, and so we weight by 1/2).

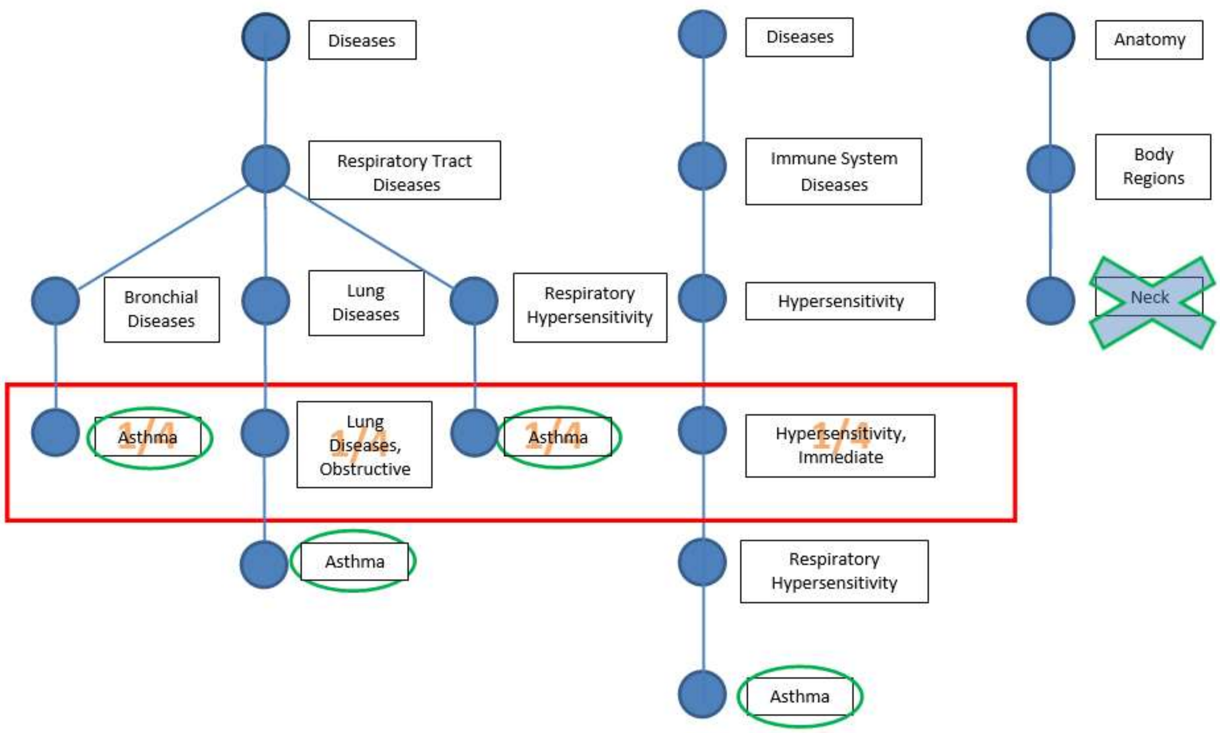

Figure C.1: Process used to aggregate MeSH terms in order to construct fields. Example shows MeSH terms Asthma and Neck.

Figure C.2 plots the distribution of 4-digit MeSH terms per article for the six different time periods. The number of MeSH terms assigned to articles increases over time.

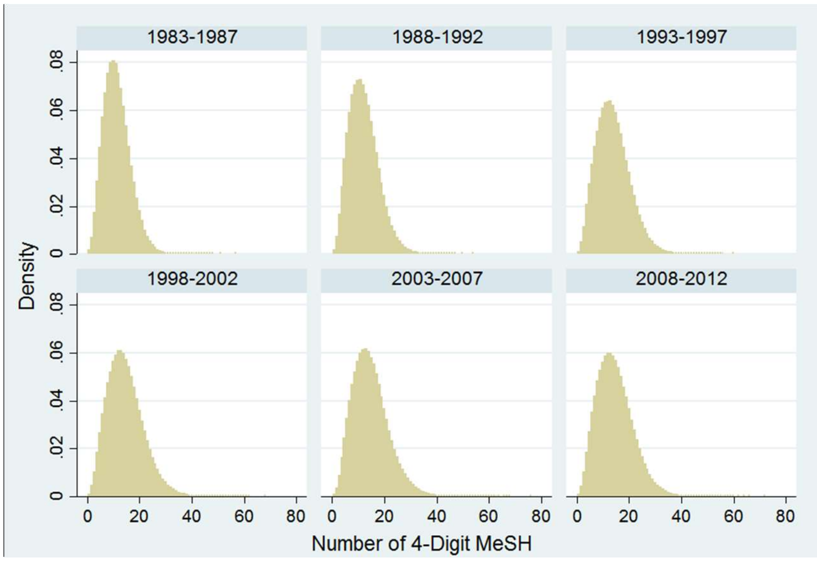

Figure C.2: Distribution of the number of 4-digit MeSH terms per article.

Figure C.3 shows the mean number of MeSH4 terms as well as selected quantiles (the 10th, 25th, 50th, 75th, 90th, and 99th quantiles) of the distribution of the number of terms. The mean and median number of MeSH4 terms increases from roughly 10 per article in the 1970s to roughly 12 in 1974 after which the number of MeSH4 terms dips before gradually increasing to roughly 12 after 2000. In addition to the upward trend there is an increase in dispersion of the number of MeSH4 terms, with no increase in the 10th and 25th quantiles in the most recent years.

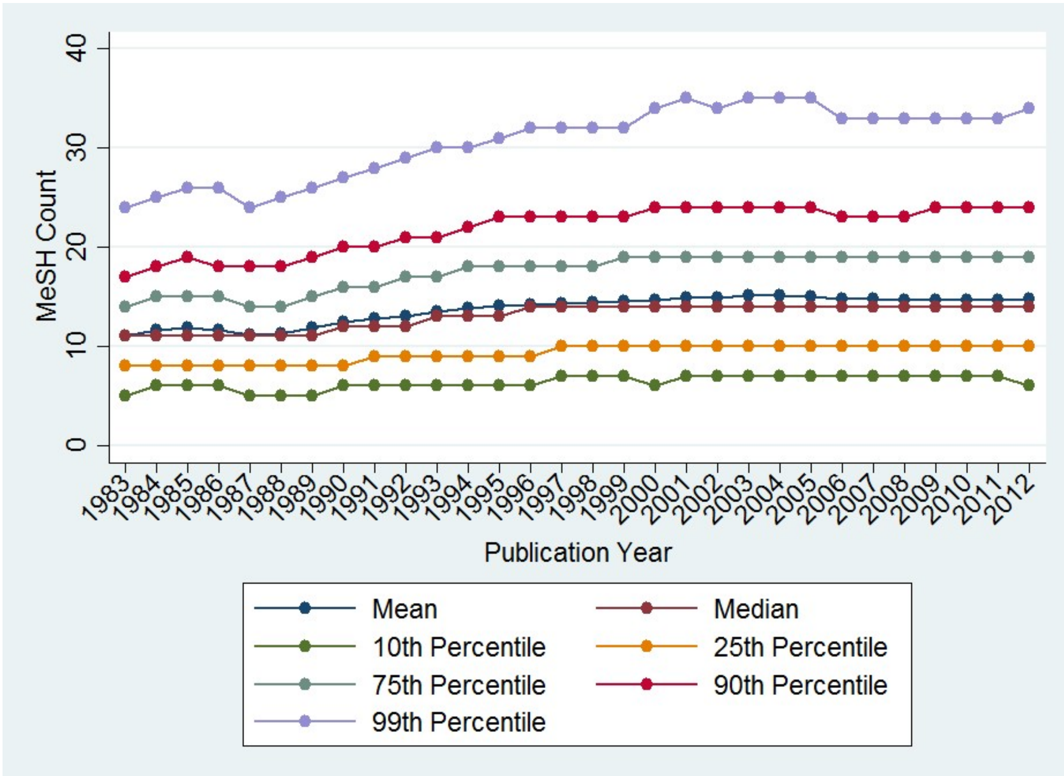

Figure C.3: Quantiles of the distribution of the number of MeSH4 terms per article by publication year.

D. Formal Definition of Metrics

The below table gives the formal description of each metric defined at the field-period level. A bin *b* is defined by a field-period pair; *i* indexes articles in bin *b*; w_i is the fraction of article *i* allocated to the MeSH4 field of *b*; and $forward\ (backward)\ citation\ count_i$ is article *i*'s forward (backward) citation count.

Table D.1: Formal definition of field-period metrics

1. Radical - Generative	
Concepts	Let $\mathbb{C}$ be the set of all top concepts.[1] Let $\mathbb{A}^{pc}$ be the set of all articles published in time period p that use concept $c \in \mathbb{C}$ in the first year it is used, its "vintage year". Let $A^{pc} = \#\mathbb{A}^{pc}$ be the number of articles in the set $\mathbb{A}^{pc}$. Consider article $a \in \mathbb{A}^{pc}$. Define α_{af} as the fraction of article a that belongs to field f. Then the fraction of concept c that belongs to field f is given by $\gamma_{fpc} = \frac{1}{A^{pc}} \sum_{a \in \mathbb{A}^{pc}} \alpha_{af}$. Finally, the total number of top concepts that originate in field f in time period p is given by $concepts_{fp} = \sum_{c \in \mathbb{C}} \gamma_{fpc}$.
BMentN	Let $\mathbb{C}$ be the set of all top concepts. Let $\mathbb{A}^{pcN}$ be the set of all articles published in time period p that use concept $c \in \mathbb{C}$ within N years of its vintage year. Consider article $a \in \mathbb{A}^{pcN}$. Define α_{af} as the fraction of article a that belongs to field f. Then the total number times that field f uses concept c within N years of its vintage is given by $n_{fpNc} = \sum_{a \in \mathbb{A}^{pcN}} \alpha_{af}$. Finally, the total number of top concept mentions that belong to field f in time period p is $BMentN_{fpN} = \sum_{c \in \mathbb{C}} n_{fpNc}$.
2. Radical - Destructive	
BCiteAge	Consider an article *j* that is cited by article *i*. The age of the citation to article *j* is $\text{backward citation age}_{ij} = i'\text{s publication year} - j'\text{s publication year}$ and article *i*'s average backward citation age is $\text{average backward citation age}_i = \frac{1}{M_i} \sum_j backward\ citation\ age_{ij}$ where M_i is the number of backward citations for paper *i*.[2] Then the backward citation age among articles in bin *b* (i.e. some field period pair) is $\text{BCiteAge}_b = \frac{1}{\sum_{i \in b} w_i} \sum_{i \in b} w_i \times (average\ backward\ citation\ age_i)$, where w_i gives the share of weight on article *i* among all articles in bin *b.*
3. Risky	
FCiteVar	Suppose there are K_b articles in bin b, $\text{FCiteVar}_b = \frac{1}{K_b} \sum_{i \in b} (forward\ citation\ count_i - FCiteMean_b)^2$.
4. Multidisciplinary	
BHerfCite	Let π_{if} be the set of articles in field *f* that article *i* cites. Then the weighted number of articles from field *f* that article *i* cites is $\sum_{k \in \pi_{if}} w_k$, where w_i gives the share of weight on article *i* among all articles in field *f.*

	Each article contributes to the backward citation count in bin *b* according to its bin *b* weight. Thus the number of citations from all articles in bin *b* to articles in field *f* is $$R_b^f = \sum_{i \in b} w_i \sum_{k \in \pi_{if}} w_k.$$ Summing R_b^f over all cited fields gives the total number of citations from articles in bin *b*, $$R_{b=} \sum_{i \in b} w_i \sum_f \sum_{k \in \pi_{if}} w_k.$$ Let $\omega_b^f = \frac{R_b^f}{R_b}$ be the share of *b*'s backward citations to field *f*. Then $$\text{BHerfCite}_b = 1 - \sum_f \left(\omega_b^f\right)^2.$$
BHerfMentN	Let $\mathbb{C}$ be the set of all top concepts.[1] Let $\mathbb{A}^{pcN}$ be the set of all articles published in time period p that use concept $c \in \mathbb{C}$ within N years of its vintage year. Consider article $a \in \mathbb{A}^{pcN}$. Define α_{af} as the fraction of article a that belongs to field f. Then the total number of times that field f in period p uses concept c within N years of its vintage is given by $n_{fpNc} = \sum_{a \in \mathbb{A}^{pcN}} \alpha_{af}$. The total number of that field f in period p uses *any* concept within N years of its vintage is given by $m_{fpN} = \sum_{c \in \mathbb{C}} n_{fpNc}$. The Herfindahl for field f in period p is given by $herf_{fpN} = 1 - \sum_{c \in \mathbb{C}} \left({}^{n_{fpNc}}/_{m_{fpN}} \right)^2$.
5. Wide Impact	
FHerfCite	Let θ_{if} be the set of articles in field *f* that cite article *i*. Then the number of articles from field *f* that cite article *i* is $$\sum_{j \in \theta_{if}} w_j.$$ Each cited article contributes to the total forward citation count in bin *b* according to its bin *b* weight. Thus the number of citations from field *f* to all articles in bin *b* is $$C_b^f = \sum_{i \in b} w_i \sum_{j \in \theta_{if}} w_j.$$ Summing C_b^f over all citing fields gives the total number of citations to articles in bin *b*, $$C_{b=} \sum_{i \in b} w_i \sum_f \sum_{j \in \theta_{if}} w_j.$$ Let $s_b^f = \frac{C_b^f}{C_b}$ be the share of *b*'s forward citations in field *f*. Then $$\text{FHerfCite}_b = 1 - \sum_f \left(s_b^f\right)^2.$$
FHerfMent	Let $\mathbb{C}^v$ be the set of all top concepts with a vintage v. Let $\mathbb{A}^c$ be the set of articles that use concept $c \in \mathbb{C}^v$ in the vintage year v. Consider an article $a \in \mathbb{A}^c$. Define α_{af} as the fraction of article a that belongs to field f. Then the fraction of concept c that belongs to field f is $\gamma_{cf} = \frac{1}{A^c} \sum_{a \in \mathbb{A}^c} \alpha_{af}$.

	Let $\mathbb{A}_{\infty}^{c}$ be the set of articles that *ever* use concept $c \in \mathbb{C}^{v}$. Then the number of times that field g ever mentions concept c is given by is $n_{cg\infty} = \sum_{a \in \mathbb{A}_{\infty}^{c}} \alpha_{ag}$. The Herfindahl index is given by $$herf_{fv} = 1 - \sum_{g \in \mathbb{F}} \sum_{c \in \mathbb{C}^{v}} \left\{ \left[\frac{\gamma_{cf} \times n_{cg\infty}}{\sum_{g \in \mathbb{F}} \sum_{c \in \mathbb{C}^{v}} (\gamma_{cf} \times n_{cg\infty})} \right]^{2} \right\}$$
6. Growing Impact	
FCiteAge	Consider an article *i* that is cited by article *j*. The age of the citation from article *j* is $$forward\ citation\ age_{ij} = j's\ publication\ year - i's\ publication\ year$$ and article *i*'s average forward citation age is $$average\ forward\ citation\ age_{i} = \frac{1}{N_i} \sum_{j} forward\ citation\ age_{ij}$$ where N_i is the number of forward citations for paper *i*.[2] $$\text{FCiteAge}_{b} = \frac{1}{\sum_{i \in b} w_i} \sum_{i \in b} w_i \times (average\ forward\ citation\ age_{i}).$$
7. High Impact	
FCiteMean	$$\text{FCiteMean}_{b} = \frac{1}{\sum_{i \in b} w_i} \sum_{i \in b} w_i \times (forward\ citation\ count_{i}).$$
FCiteN	Order the K_b articles in bin *b* by forward citation count. Index the ordered articles by *j* = 1, 2, ..., K_b where *j* = 1 corresponds to the article with the highest forward citation count, *j* = 2 corresponds to the second highest citation count, and so on. Then compute, $$\text{FCite}N_{b} = forward\ citation\ count_{z} \text{ where } z = \underset{j}{\text{argmin}} \left\{ j, satisfies \sum_{i=1}^{j} w_i \geq (1 - \frac{N}{100}) \sum_{k \in b} w_k \right\}.$$

Notes:
1. Top Concepts: Top concepts defined as one of the 10,128 most highly used 1-, 2-, or 3-grams that first appeared in a MEDLINE title or abstract between 1983 and 2012.
2. FCiteAge: *forward citation age*$_{ij}$ is set to missing if either the value is negative or if *j*'s publication year is not valid. N_i is the number of forward citations to article *i* that have non-missing ages.