# 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** |
| **Radical Generative** | | | |
| 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 n-grams  Mentions (*BMentT*)—Use of important new n-grams | | | |
| **Radical Destructive** | | | |
| 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 | | | |
| **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 | | | |
| **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 (*BHerfMentT*)—Breadth of important n-grams used | | | |
| **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 n-grams introduced | | | |
| **Growing Impact** | | | |
| Often fragile in their early stages |  |  |  |
| **Metrics of time path of utilization of knowledge:**  Forward Citation Age (*FCiteAge*)—Age of citations | | | |
| **Impact** | | | |
| Scientific understanding advances dramatically | Profoundly impact … 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. 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: https://mbr.nlm.nih.gov/Download/2009/WordCounts/wrd\_stop
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.
9. Stem each word from each n-gram using the module *Lingua::Stem* from the Comprehensive Perl Archive Network (CPAN).
10. 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 Clarivate Analytics’ the Science Citation Index ExpandedTM 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 is 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.

Consider the MeSH term “Asthma” This term has four separate identifiers: “C08.127.108”, “C08.381.495.108”, “C08.674.095”, and “C20.543.480.680.095”. Thus, “Asthma” falls under the first level MeSH term “Diseases” (identified by “C”). It also falls under the second-level terms “Respiratory Tract Diseases” (“C08”) and “Immune System Diseases” (“C20”). The problem arises because MEDLINE records contain only the MeSH terms themselves, not their identifiers. For instance, if a MEDLINE record is tagged with the MeSH term “Asthma”, we would not know whether it was the “Asthma” that is beneath “Respiratory Tract Diseases” or “Immune System Diseases”.

If we wanted to aggregate all MeSH terms to the second level, we would have to find a way to split “Asthma” between “Respiratory Tract Diseases” and “Immune System Diseases”. 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, “Asthma” would fall under the fourth-level terms

* Lung Diseases, Obstructive [C08.381.495]
* Hypersensitivity, Immediate [C20.543.480]
* Asthma [C08.127.108]
* Asthma [C08.674.095]

In this case, we assign a quarter of the raw term “Asthma” to each of these four fourth-level terms. Thus, overall, 1/4 will be assigned to “Lung Diseases, Obstructive”, 1/4 to “Hypersensitivity, Immediate”, and 1/4+1/4=1/2 to “Asthma” itself.

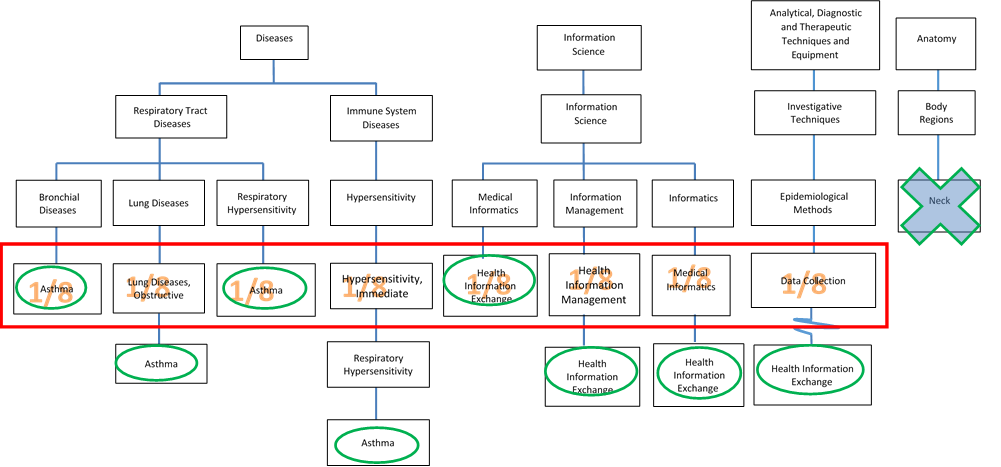
Our last step is to apportion each article indexed in MEDLINE to the newly aggregated MeSH terms. Again, examples are the most illustrative. Suppose we come across an article that is tagged by “Asthma” and we want to aggregate to the fourth level. We know from above that 1/4 will be assigned to “Lung Diseases, Obstructive”, 1/4 to “Hypersensitivity, Immediate”, and 1/4+1/4=1/2 to “Asthma” 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 terms “Neck” (identified by A01.598) and “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). Fig C.1 illustrates this example. By the process discussed above, 1/4 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). Since the lowest level of aggregation for “Neck” is the third level, we cannot assign it to a fourth-level term. We deal with this by simply eliminating “Neck”. It is too highly aggregated for our purposes.

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

* 1/2\*1/4=1/8 to “Lung Diseases, Obstructive”
* 1/2\*1/4=1/8 to “Hypersensitivity, Immediate”
* 1/2\*1/4=1/8 to “Asthma”
* 1/2\*1/4=1/8 to “Asthma”
* 1/2\*1/4=1/8 to “Health Information Exchange”
* 1/2\*1/4=1/8 to “Health Information Management”
* 1/2\*1/4=1/8 to “Medical Informatics”
* 1/2\*1/4=1/8 to “Data Collection”

Obviously, 1/8+1/8+1/8+1/8+1/8+1/8+1/8+1/8=1. Thus, the article that was originally tagged by the three MeSH terms “Asthma”, “Neck” and “Health Information Exchange” is now apportioned between eight 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 “Asthma” equally across “Lung Diseases, Obstructive”, “Hypersensitivity, Immediate”, “Asthma”, and “Asthma”). Second, we weight this apportionment by the inverse of the number of original MeSH terms of the proper level that tag the article (e.g. our hypothetical article was tagged by three original MeSH terms, but only two at the proper level of aggregation, and so we weight by 1/2).



**Fig C.1. Process used to aggregate MeSH terms in order to construct fields.** Example shows MeSH terms “Asthma” and “Neck.”

Fig 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.

 C:\Users\JoeS\Downloads\meshcountdist.tif

**Fig C.2.** **Distribution of the number of 4-digit MeSH terms per article.**

Fig 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.

C:\Users\JoeS\Downloads\meshcountmoments.tif

**Fig C.3. Quantiles of the distribution of the number of MeSH4 terms per article by publication year.**

## D. Formal Definition of Metrics

Common notation for the metric definitions include:

* denotes a field. When it is necessary to reference two fields simultaneously, can also denote a field.
* denotes a time period.
* denotes a set of articles. We frequently use the set of articles published in field-period pair , which we denote .
* denotes the set of top n-grams (concepts). These are the top 0.01 percent of n-grams in terms of appearances in the titles and abstracts of MEDLINE from each vintage (including all tied n-grams in the case of ties at the threshold) — a total of 10,229 top n-grams (including 589 due to ties) with vintages between 1983 and 2012. We use these top n-grams to construct our text-based metrics.
* *N* denotes a count. For example, denotes the number of articles in field-period pair *fp*
* We denote weights by . For instance, denotes the fraction of article that belongs to field and denotes article ’s share of all articles in field and period *p*.
* When identifying fields in which articles are cited, indicates the field that is cited and denotes the field that makes a citation.

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

|  |  |
| --- | --- |
| 1. **Radical - Generative** | |
| Concepts | Let be a top n-gram. Let  be an article, published in period , that uses in ’s vintage (birth) year. For instance, if ’s vintage is 2004, and , then is published in 2004 and uses . Note that, in this example, corresponds to the period 2003-2007.  Since is the fraction of article that belongs to field , the total number of normalized articles that belong to field are published in period , and use n-gram in ’s vintage (birth) year is .  Since is apportioned to a field in the same way as , the total number of top n-grams that originate in field during period is defined as:  Note that is normalized by to ensure that field-period pairs with many articles that use in its vintage year do not get counted as using more n-grams. |
| BMentT | Let be a top n-gram. Let be an article, published in period , that uses within *T* years of ’s vintage (birth) year. For instance, if ’s vintage is 1999, , and , then is published any time between 1999-2001 and uses . Note that, in this example, corresponds to the period 1998-2002.  Since is the fraction of article that belongs to field , the total number of (weighted) articles that belong to field are published in period , and use n-gram within years of ’s vintage is .  Since is apportioned to a field in the same way as , the total number of top n-grams that are used within years of vintage and belong to field-period pair is defined as: |
| 1. **Radical - Destructive** | |
| BCiteAge | Consider an article *j* that is cited by article *i.* The age of the citation to article *j* is  and article *i*'s average backward citation age is  where is the number of backward citations for paper *i*. Then the backward citation age among articles in field period pair *fp* is |
| 1. **Risky** | |
| FCiteVar | Let denote the count of forward citations to article *i* and , that is, the unfractionalized mean of forward citations in field period pair *fp.* |
|  |  |
| 1. **Multidisciplinary** | |
| BHerfCite | Let be the set of articles in field that article *i* references. Then the weighted number of articles from field that article *i* cites is .  Each article contributes to the backward citation count in field period pair *fp* according to its weight, . Thus, the number of citations from all articles in field period pair *fp* to articles in cited field is  Summing over all fields gives the total number of backward citations from articles in field period pair *fp*:  Let be the share of field period pair *fp*’s backward citations to field . Then |
| BHerfMentT | Let be a top n-gram. Let be an article, published in period , that uses within years of ’s vintage (birth) year. For instance, if ’s vintage is 1989, , and , then is published any time between 1989-1998 and uses . Note that, in this example, corresponds to the period 1988-1992.  Since is the fraction of article that belongs to field , the total number of articles that belong to field , are published in period , and use n-gram  within years of ’s vintage is .  Since is apportioned to a field in the same way as , the total number of n-grams that are used within  years of vintage and belong to field-period pair  is given by .  The fraction of n-gram uses within years of vintage that belongs to field-period pair is then given by:  The backward-looking Herfindahl index for fie in period using n-grams within years of vintage is defined as: |
| 1. **Wide Impact** | |
| FHerfCite | Let be the set of articles in field that cite article *i*. Then the weighted number of articles from field that cite article *i* is  Each cited article contributes to the total forward citation count in field period pair *fp* according to its weight . Thus the number of citations from field to all articles in field period pair *fp* is  Summing over all fields gives the total number of citations to articles in field period pair *fp*:  Let be the share of field period pair *fp*’s forward citations in field . Then |
| FHerfMent | Let be a top n-gram. Let  be an article, published in period , that uses in ’s vintage (birth) year. Let be an article, published in *any* period, that uses .  The fraction of originated in field is . The total number of n-gram  mentions by field (in any period) is .  The fraction of all weighted n-gram mentions by field is given by:  The forward-looking Herfindahl index for field in period is defined as: |
| 1. **Growing Impact** | |
| FCiteAge | Consider an article *i* that is cited by article *j*. The age of the citation from article *j* is  and article *i*'s average forward citation age is  where is the number of forward citations for paper *i*.2 |
| 1. **Impact** | |
| FCiteMean |  |
| FCiteP | Order the articles in field period pair *fp* by forward citation count. Index the ordered articles by *j* = 1, 2, ..., 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  ,  with *P* denoting the *P*th percentile of the citation distribution. |
| Notes:  1. Top Concepts: A top concept is 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.   1. 2. FCiteAge: *forward citation ageij* is set to missing if either the value is negative or if *j*’s publication year is not valid. is the number of forward citations to article *i* that have non-missing ages. | |