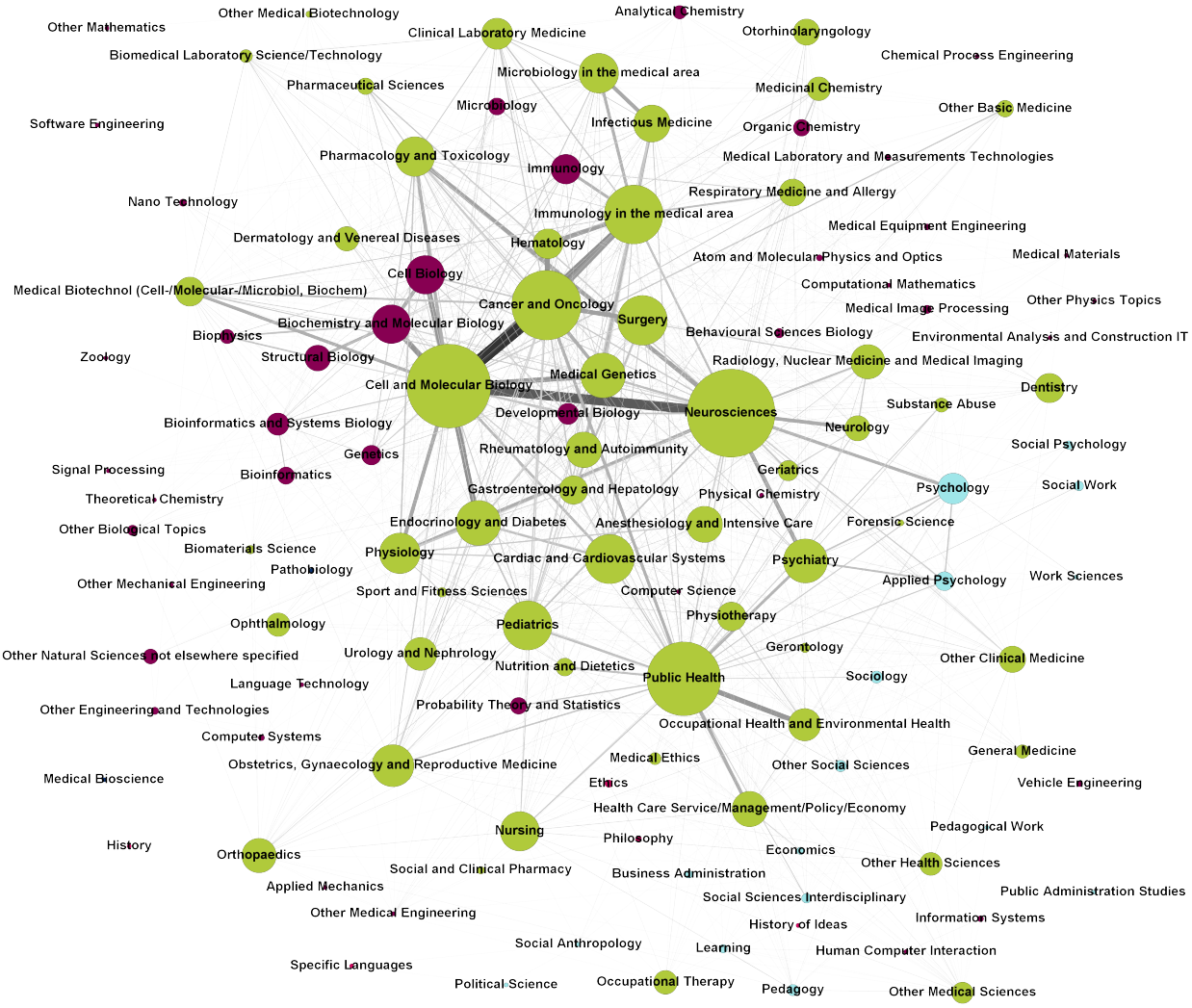

BIBLIOMETRIC HANDBOOK FOR KAROLINSKA INSTITUTET

UNIVERSITY LIBRARY BIBLIOMETRIC TEAM 2014

CATHARINA REHN, CARL GORNITZKI, AGNE LARSSON & DANIEL WADSKOG

Karolinska Institutet

TABLE OF CONTENTS

1. INTRODUCTION

1.1 ABOUT THE HANDBOOK

The Bibliometric Handbook for Karolinska Institutet was first published in 2005 with the intention to enhance the understanding of bibliometrics by describing bibliometric methodology in general and, more specific, how it is used at Karolinska Institutet. This version was updated and published in 2014.

Although the main purpose of the handbook is to provide transparency for Karolinska Institutet employees and affiliates into the bibliometric analyses made within their own organization, the handbook is hopefully useful for anyone – for example researchers, policy makers and management at different levels – interested in learning more about bibliometrics.

1.2 WHAT IS BIBLIOMETRICS?

One of the earliest definitions of bibliometrics describes it as "the application of statistical and mathematical methods to books and other media of communication" (Pritchard 1969).

Today, bibliometrics is often used to assess scientific research through quantitative studies on research publications. Bibliometric analyses are based on the assumption that most scientific discoveries and research results eventually are published in international scientific journals where they can be read and cited by other researchers.

Evaluative bibliometrics – "quantitative measurements of qualitative aspects (such as 'quality' or 'reputation') of the science system" (van Leeuwen, 2004) – is based on the assumption that the number of citations to a journal article can be considered to reflect the article's impact on the scientific community.

The term *bibliometric indicators* is often used for the results of a bibliometric analysis. One of the definitions of the term *indicator* in the Oxford English Dictionary is "That which serves to indicate or give a suggestion *of* something; an indication *of*" (*Oxford English dictionary* 2000). This draws attention to the fact that the results describe a reality that is too complex to be measured merely by statistics or numbers.

In a glossary produced by the United Nations Development Programme Evaluation Office, there is another definition that seems close to how the word indicator is used in bibliometrics (2002, p. 101):

> "Indicator: Signal that reveals progress (or lack thereof) towards objectives; means of measuring what actually happens against what has been planned in terms of quantity, quality and timeliness. An indicator is a quantitative or qualitative variable that provides a simple and reliable basis for assessing achievement, change or performance."

1.3 ANALYZING BIBLIOGRAPHIC REFERENCES

This section contains several examples of the kind of results you can get from a statistical analysis of standard bibliographic references like the one below:

Annu Rev Med. 2006;57:119-37.
Pharmacogenomics and individualized drug therapy.

Eichelbaum M, Ingelman-Sundberg M, Evans WE.

Dr Margarete Fischer Bosch Inst Clin Pharmacol, Stuttgart, D-70376 Germany
Karolinska Inst, Div Mol Toxicol, IMM, Stockholm, SE-17177 Sweden
St Jude Childrens Hosp, Memphis, TN 38105 USA

Pharmacogenetics deals with inherited differences in the response to drugs. The best-recognized examples are genetic polymorphisms of drug-metabolizing enzymes, which affect about 30% of all drugs. Loss of function of thiopurine S-methyltransferase (TPMT) results in severe and life-threatening hematopoietic toxicity if patients receive standard doses of mercaptopurine and azathioprine. Gene duplication of cytochrome P4502D6 (CYP2D6), which metabolizes many antidepressants, has been identified as a mechanism of poor response in the treatment of depression. There is also a growing list of genetic polymorphisms in drug targets that have been shown to influence drug response. A major limitation that has heretofore moderated the use of pharmacogenetic testing in the clinical setting is the lack of prospective clinical trials demonstrating that such testing can improve the benefit/risk ratio of drug therapy.

MeSH Terms:
Biotransformation/genetics
Cytochrome P-450 Enzyme System/genetics
Glucuronosyltransferase/genetics
Humans
Methyltransferases/genetics
Polymorphism, Genetic/genetics
Receptors, Adrenergic, beta-2/genetics
Sodium Channels/genetics

PUBLICATION YEAR

An analysis of publication years can for example show trends in how much a unit publishes compared to the rest of the world, or to similar units.

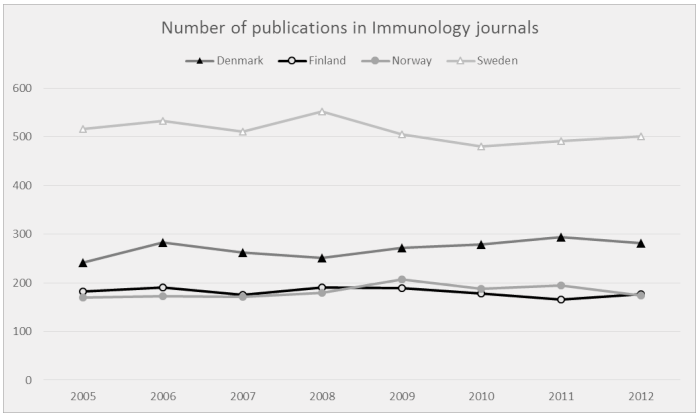

Figure 1. The number of articles from Sweden, Denmark, Norway and Finland in immunology journals 2005–2012.

JOURNAL TITLE

An analysis of journal titles can for instance give an overview of the publication pattern of a certain unit.

Journal	*Publications*
JOURNAL OF IMMUNOLOGY	75
ALLERGY	74
VACCINE	60
BONE MARROW TRANSPLANTATION	50
JOURNAL OF INFECTIOUS DISEASES	36

Table 1. The five most frequent immunology journals used for publication 2008–2012 by authors at Karolinska Institutet.

AUTHOR NAMES

It may be of interest to identify prolific authors in a specific country or at a specific unit.

Author	*Publications*
Blom, AM	41
Ringden, O	36
Wickman, M	31
Ljungman, P	30
Wahren, B	28
Mattsson, J	28
Hammarstrom, L	28
van Hage, M	28
Cardell, LO	27
Riesbeck, K	24

Table 2. The ten most prolific Swedish journal article author names within the field of immunology 2008–2012.

Additional information can also be had by analyzing the co-publication patterns of authors in a specific area.

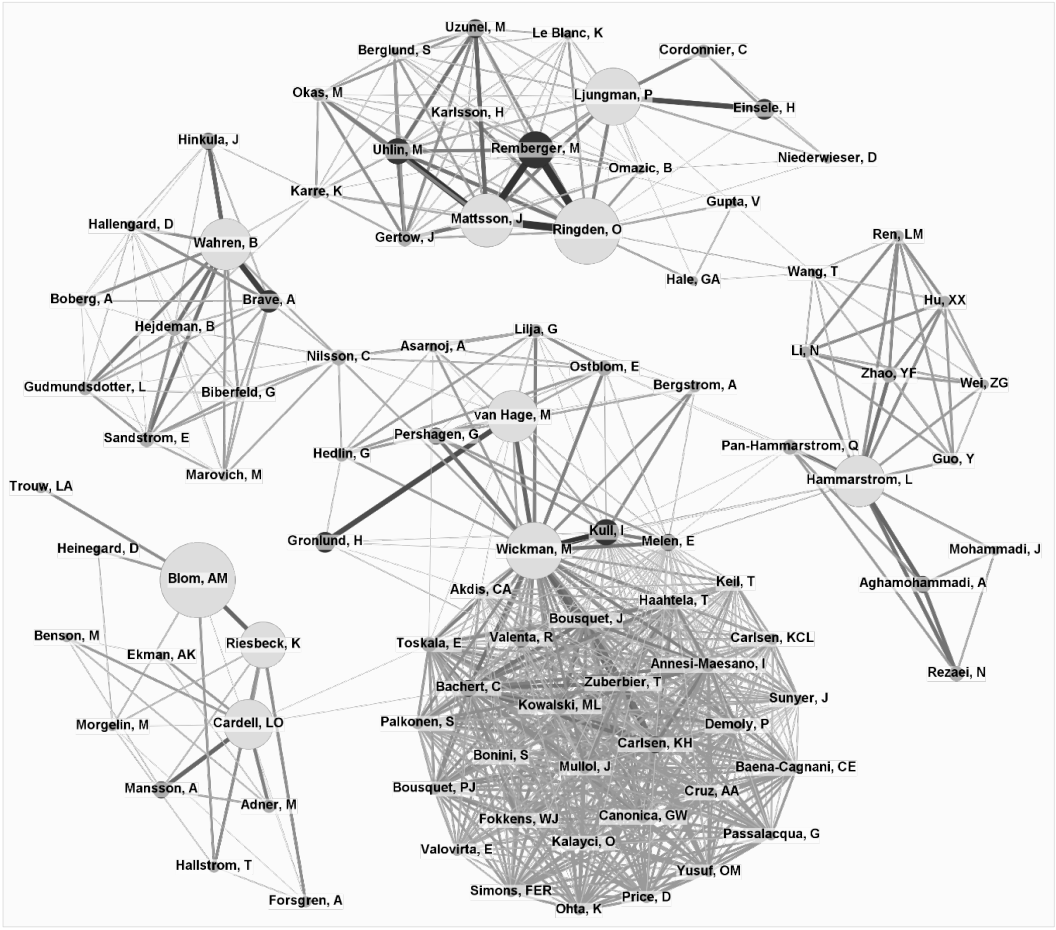

Figure 2. The co-publication pattern of the ten most prolific Swedish journal article authors that publish in immunology journals (2008–2012). Each author included in the analysis has at least 5 copublications with any of the ten most prolific authors.

AUTHOR ADDRESSES

Analogous to an analysis of author names, an analysis of author addresses can identify prolific countries, universities or other organizations and give an overview of the co-publication patterns.

Country	*Publications*
Sweden	501
Denmark	281
Finland	177
Norway	174
Iceland	10

Table 3. The number of immunology publications in 2012 from the five Nordic countries.

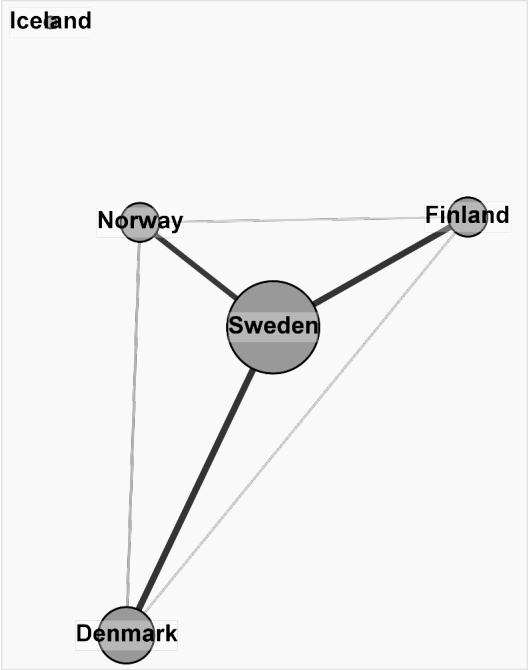

Figure 3. Co-publication patterns in immunology between the Nordic countries in 2012 based on publications in immunology journals.

KEYWORDS

If you have adequate keywords assigned to the publications it is possible to study which subjects appear often in the publications of a unit and connections between them, a so called co-word analysis. This kind of analysis is easier to make if the reference contains terms from a controlled vocabulary like MeSH (Medical Subject Headings from the National Library of Medicine, USA).

MeSH	*Publications*
Arthritis, Rheumatoid	184
Antirheumatic Agents	55
Tumor Necrosis Factor-alpha	30
Immunoglobulin G	24
Antibodies, Monoclonal	23
Autoantibodies	23
Genetic Predisposition to Disease	23
Receptors, Tumor Necrosis Factor	23
Arthritis, Experimental	21
Lupus Erythematosus, Systemic	17

Table 4. The ten most frequent keywords in publications by one specific author.

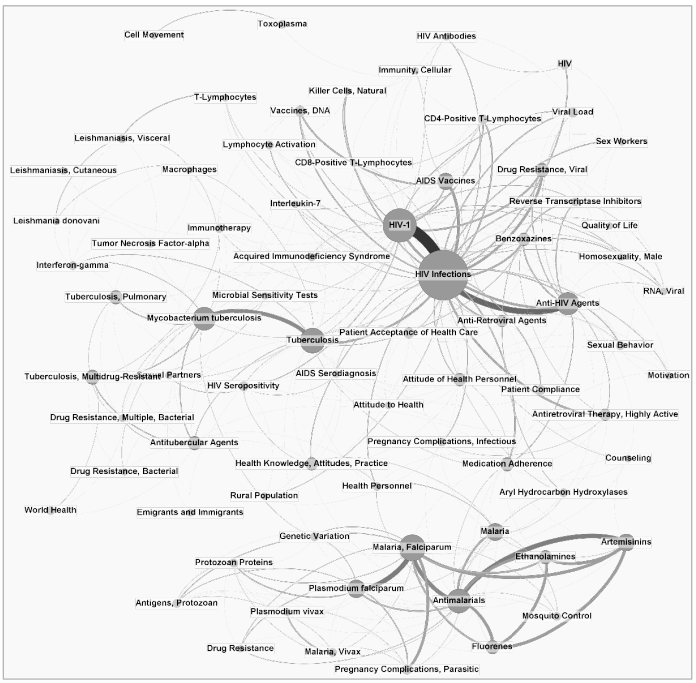

Figure 4. The co-occurrence pattern of MeSH-terms in publications by authors at one of the Karolinska Institutet research centers. (To be included in this picture, a MeSH term needs at least three occurrences in the analyzed set of publications.)

COMBINATIONS OF THE ABOVE

The analyses above are often improved by combining more than one aspect.

By combining author names or author addresses and keywords it is possible to localize prolific authors or organizations in a specific field. This also makes it possible to identify authors that are connected through common subjects – something that may identify existing research networks or the possibility of a new research network.

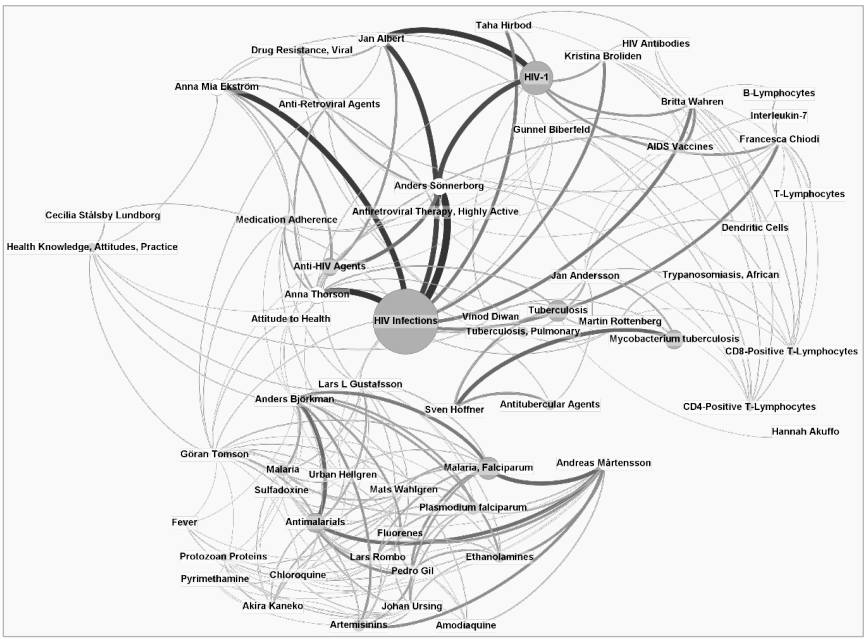

Figure 5. Connections through subjects (represented by MeSH terms) between authors at one of the Karolinska Institutet research centers.

1.4 ANALYZING REFERENCE LISTS AND CITATIONS

When bibliographic records include reference lists to cited articles, we can extend our statistical analysis on the connections between various publications. In general, the additional information obtained from reference lists can supply at least two more aspects to a bibliometric study:

- The possibility to find publications in the same area by identifying the publications that cite (refer to) or are cited by the publications that you already have identified.
- The possibility of a bibliometric quality assessment.

BIBLIOGRAPHIC COUPLING

A bibliographic coupling analysis connects publications that share items in their reference lists (see figure 6), that is, refer to the same publications.

The assumption behind bibliographic coupling is that publications within the same subject share core material and the more like the publications, the more like the reference lists. One specific trait of this method is that it makes it possible to find conceptual connections between articles that are so new that they haven't had the time to be cited yet.

An example on how bibliographic coupling can be used is the case where you find a highly relevant article that refers to an important previous article. You may then do a bibliometric search to see what other articles that refers to this previous article to see if there are newer relevant articles on the same subject as the first article.

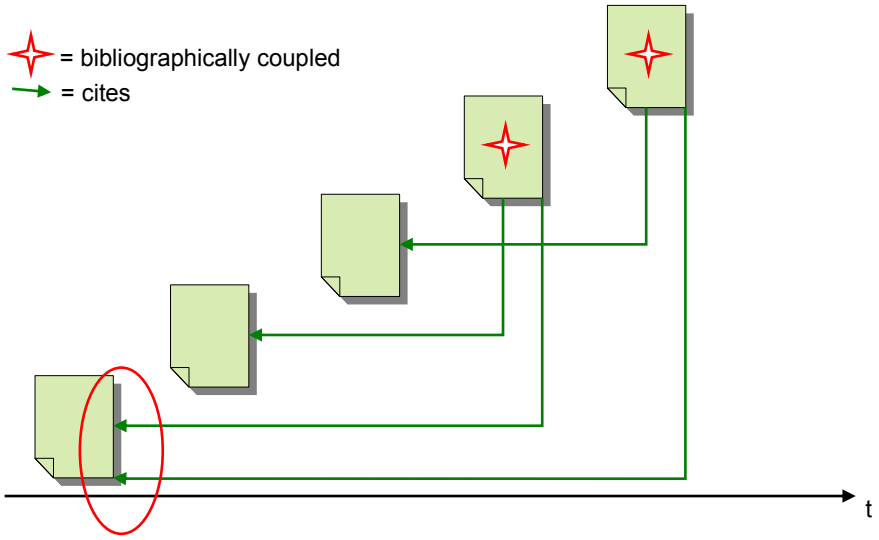

Figure 6. Bibliographic coupling analyzed via shared references. The more recent star-marked articles may be subject-related to each other, since they refer to the same older article.

CO-CITATION ANALYSIS

A co-citation analysis studies reference-pairs, i.e. papers cited (referred to) by the same publication (see figure 7). By doing a co-citation analysis you may find older articles that are related to each other, even though they don't refer to each other.

This type of analysis will usually generate clusters with highly cited articles since two highly cited papers are more likely to be co-cited in several reference lists than two lowly cited ones.

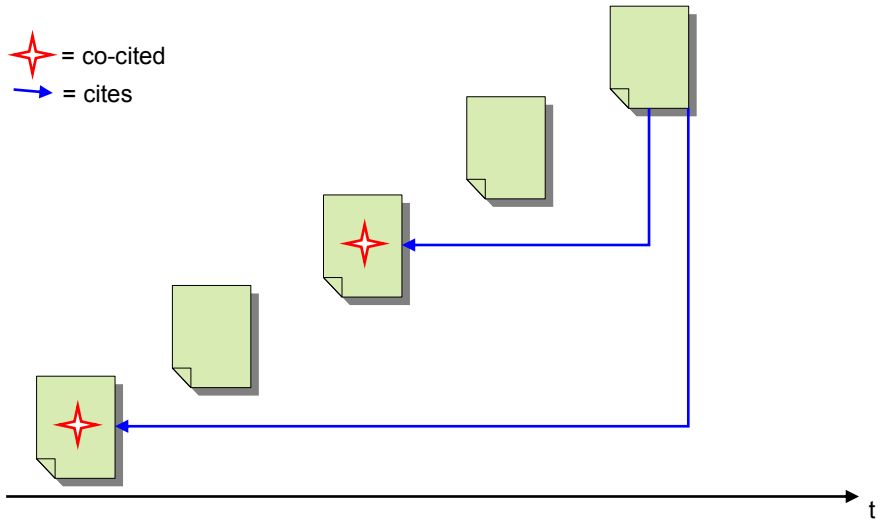

Figure 7. Co-citation analyzed via shared reference lists. The star-marked articles may have something in common, since a more recent article refers to them both.

QUALITY ASSESSMENT

If citation rates of a unit's publications are high compared to the expected citation rate per publication this shows that the unit's articles have had significant impact on the scientific society, which in turn may indicate that the research is of high quality.

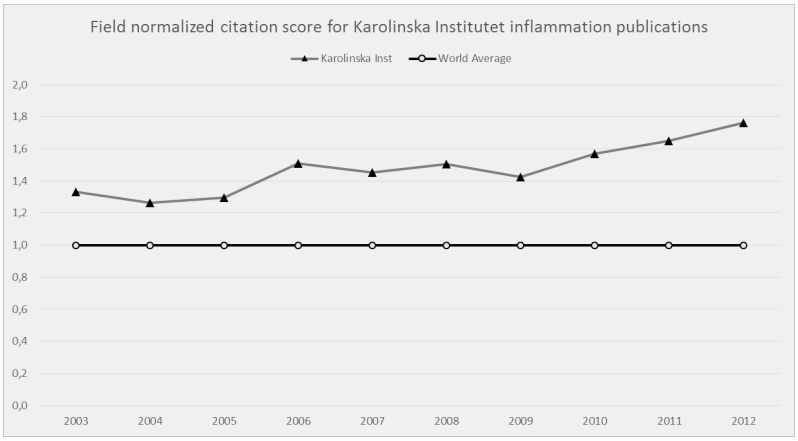

Figure 8. The mean value of the number of citations to Karolinska Institutet articles in immunology journals in relation to the world average of citations to articles within the same research area. We can see a positive trend, which may be an indication of improved article quality during recent years.

1.5 HOW DO YOU GET CITED?

A comprehensive summary of research on why authors cite each other has been written by Henk Moed (2005). It has been found that it is reasonable to assume that most citations are ”positive”, that is to say a sign of the fact that the citing author finds something useful in the material he cites.

Deviating citation patterns, such as negative citations, can affect an analysis of an individual article or author, but this adverse effect tends to disappear in an analysis of larger aggregations of authors, such as departments, universities or countries.

The number of citations to a publication is affected by (Glänzel, 2003, p. 61):

- *The subject matter*, and within the subject, the “level of abstraction”. The publication activity in theoretical fields (e.g., mathematics) and in engineering is lower than in experimental fields or in the life sciences. Articles in a research field with the custom to write long reference lists also receive on average more citations.

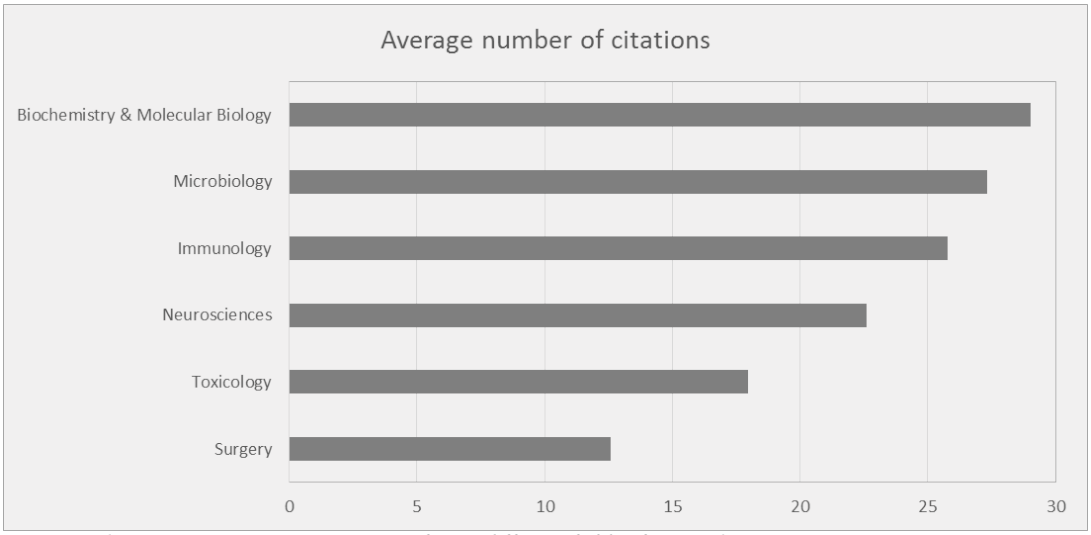

Figure 9. The average citation rate in 2004 for six different fields of research.

- *The age of the publication*. Older publications have a longer time period during which they can receive citations. Older articles are therefore on average more cited than new ones.

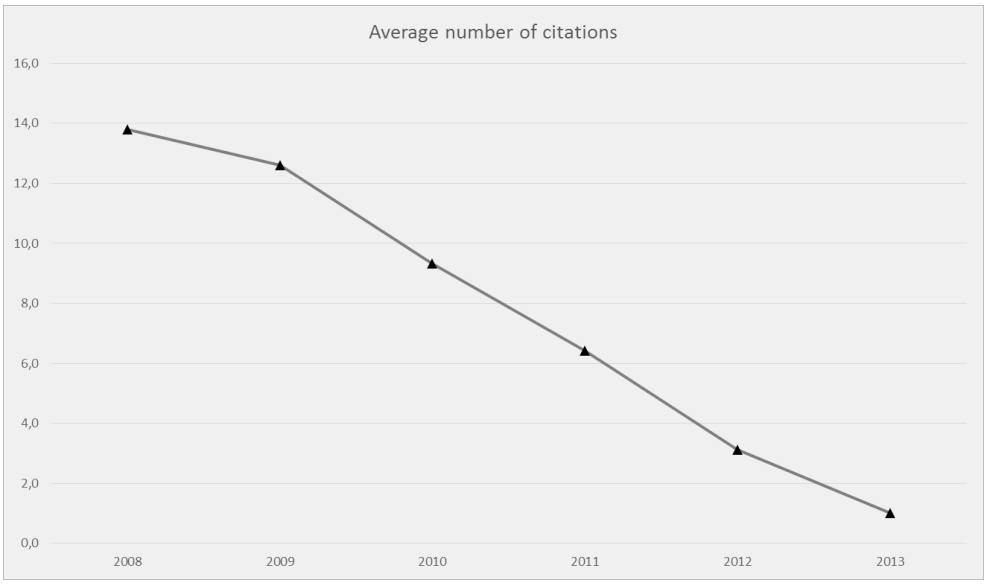

Figure 10. Average cumulative citation rates recorded in the Thomson Reuters citation indices 2014 for items published in 2008–2013 in the field of immunology.

- *The "social status"* of the publication (through the author(s) and the journal). At higher aggregation levels (e.g., at institutional or national level), the influence of the factors regarding author age and social status tends to vanish since populations at this level are rather heterogeneous.
- *The document type*. Certain types of papers, such as review papers, tend to be cited more than original articles. In most bibliometric studies only some document types are included. "Only those document types that are conveyers of relevant scientific information are taken into consideration. Such publications are, in particular, journal papers of the type research articles, letters, notes and reviews. Meeting abstracts, editorial material, corrections/errata, retractions, book reviews and other document types not listed above are only objects of special bibliometric studies." (Glänzel, 2003, p. 46)

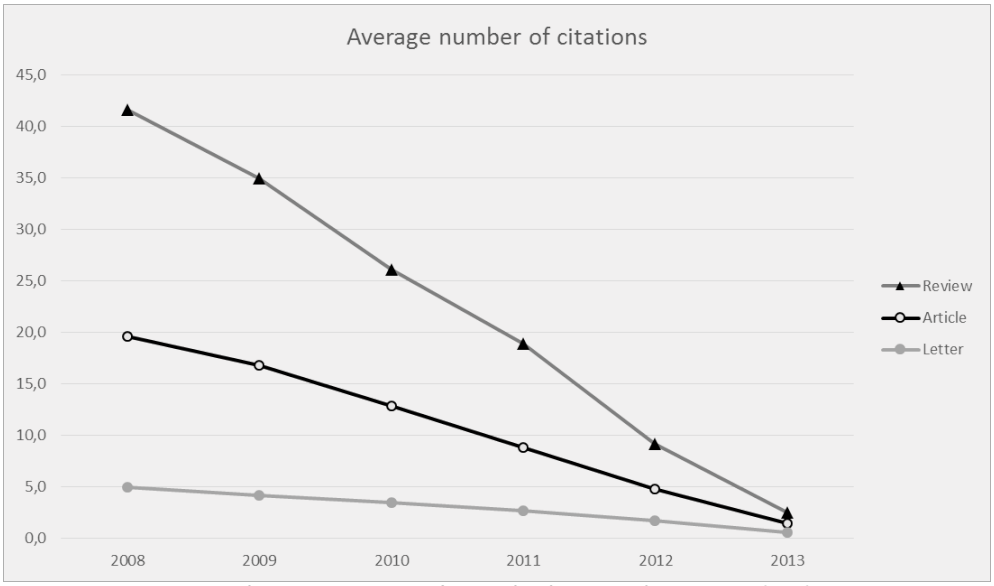

Figure 11. Average cumulative citation rates for articles, letters and reviews in the Thomson Reuters citation indices 2014 for items published in 2008–2013 in the field of immunology.

- *The length of the observation period*. Different research areas reach their ”citation peaks” at different time intervals after the publication date. It is therefore important to decide not only the publication years of the papers you wish to study (the ”time window”) but also for what years after publication you wish to count the citations (the so called ”citation window”).

In a few cases, the citation pattern differs from the one described above:

- *The Mendel effect or the “sleeping beauty”.* A single paper remains uncited for a long time until the rest of the research community discover its value and start citing it.
- *"Obliteration by incorporation".* The information in a publication has been used so much that it’s considered public knowledge and as such uncited. This effect takes very long to appear, and most bibliometric analysis only includes the latest 5 or 10 years.

However, both these varieties are very rare and do not usually affect an analysis made on a sufficiently large number of publications.

The distribution of citations to a group of articles is nearly always skewed, even if you take the above mentioned factors into consideration. Some of the publications in a group are much more cited than others. The distribution within the same group can vary from uncited articles to articles among the 1% most highly cited in the world.

This can be seen both in leading groups and in less prominent groups and it is also true for papers in high as well as for those in low impact journals. Moed says that this may be because some papers can be considered as “flags” and others as “bricks”. The bricks lay the foundation that the flags need to stand on in order to be high quality papers, but only the flag papers get cited (Moed, 2005, p. 86).

1.6 WHAT DO BIBLIOMETRIC ANALYSES MEASURE?

Bibliometric analyses result in *indicators* of research quantity and performance. They can also provide measurements of connections between researchers and research areas through statistical analysis of co-publications and citations.

Below is a list with examples of various indicators. For more indicators and full descriptions and definitions of indicators, see the publication: *Bibliometric indicators – definitions and usage at Karolinska Institutet.*

QUANTITY INDICATORS: NUMBER OF PUBLICATIONS AND CITATIONS

Examples:

- *Number of publications and citations.* The two most basic bibliometric indicators describe the number of publications and citations attributed to a group of authors (a research group, a department, a university or a country) during a specified time period.
- *World share of publications.* The unit's number of publications in relation to the world production.
- *Number of publications in Thomson Reuters citation indices.*
- *Number of publications in top-ranked journals.*

PERFORMANCE INDICATORS: NORMALIZED CITATION COUNTS

Examples:

- *Field normalized citation score (including the "crown indicator")* measures the research impact of an analyzed unit. It relates the number of citations to a set of publications to the number of citations to international publications from the same year, in the same subject area and of the same document type.
- *Top 5%* shows how many in a set of publications that belong to the 5% most cited publications in the world from the same year, in the same subject area and of the same document type. It can be expressed either as a share or as a count.
- *Journal normalized citation score.* Expresses how much a unit's publications are cited in relation to other articles in the same journals that they are published in.

JOURNAL PERFORMANCE INDICATORS: IMPACT INDICATORS

- *The Thomson Reuters Impact Factor* for a scientific journal is a mean value that corresponds to how many times an average article published in the journal has been cited.
- *Normalized journal impact* is normalized on an individual article level and corresponds to the field normalized citation score calculated for publications in one specific journal. Each publication in the journal in question is compared to other publications from the same year, of the same document type and published in journals in the same fields.

STRUCTURAL INDICATORS: PUBLICATION AND CITATION PATTERNS

Structural indicators are for example the fields in which a unit publishes and the fields in which it is cited. Further, one can make descriptions of the cognitive structure of the unit's research field, or of co-authors and the co-author's affiliations (organizations, countries etc).

One example is the use of connection maps to illustrate how much different units publish together or how a selected number of units are connected through a common field of research.

2. BIBLIOMETRIC ANALYSIS

2.1 HOW DO YOU PERFORM A BIBLIOMETRIC ANALYSIS?

Applied bibliometrics, as it is used today, analyzes scientific articles, the units that publish them, citations to these articles and connections between articles, authors and subjects.

GETTING DATA

The most common way to get data for a bibliometric analysis is to extract the information from an already existing database containing bibliographic information. The source is often one or more of the citation indices made available by Thomson Reuters, but it can also be a locally produced database of the publications from one specific unit or indeed any database containing information about the publications that are to be included in the analysis.

THOMSON REUTERS CITATION INDICES

Most bibliometric analyses use data originating from one or more of the three citation indices supplied by Thomson Reuters. (ISI – the Institute for Scientific Information – founded by Eugene Garfield in 1958 is now a part of Thomson Reuters.)

The most important Thomson Reuters citation index for medicine, life science and the natural sciences is the Science Citation Index Expanded (SCIE). This contains references to articles from more than 8 300 scientific journals (Science Citation Index Expanded, 2014). There is also a Social Sciences Citation Index, an Arts and Humanities Citation Index and a proceedings database.

Some of the advantages of the Thomson Reuters citation indices are:

- Multidisciplinary
- Go many years back
- Contain all author addresses
- Contain citation data
- Include full journal content – not just parts
- Reasonably standardized

Including all three indices, Thomson Reuters indexes about 12 000 of an estimated number of more than 27 000 active, refereed scientific journals (Ulrichsweb, 2014).

Since Thomson Reuters use the reference lists from publication records in their own indices to select what journals to include, it is reasonable to assume that the Thomson Reuters citation indices contain the most cited and most important academic journals.

Subscribers to the Thomson Reuters citation indices can, for example, access them through the web based service Web of Science. This provides the opportunity to create lists of publications and citations attributed to researchers, research groups, departments, universities or countries. It's however not suited for more complex bibliometric analyses, including the calculation of mean values or connection mapping. For this you have to purchase or download data from the Thomson Reuters

citation indices and use other applications for the calculations. Some organization use the Thomson Reuters tool Incites which provides a standardized set of more advanced indicators. A few organizations (such as Karolinska Institutet) have downloaded data into a local database and conduct custom analyses there.

As a consequence of the strong bibliometric focus on data from the Thomson Reuters citation indices, most bibliometric indicators are reliable only in research areas where publishing in scientific journals is the main mode of communication. This is often the case in natural sciences, technology and medicine, but analyses of areas within the humanities or social sciences must apply other methods as well (Moed, 2005, p. 42).

PUBMED/MEDLINE

MEDLINE is the world's largest medical database. It is produced by the National Library of Medicine (NLM), USA and covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. MEDLINE contains bibliographic citations and author abstracts from more than 5 600 biomedical journals. The database at present contains over 20 million citations dating back to 1946. Coverage is worldwide, but most records are from English-language sources or have an English abstract. A lot of work is put into the indexing of article references with the controlled vocabulary MeSH.

PubMed is NLM's own search interface to MEDLINE and includes over 23 million references. In addition to MEDLINE, PubMed retrieves (Fact Sheet; MEDLINE, PubMed, and PMC (PubMed Central): How are they different?, 2014):

- "In-process citations, which provide records for articles before they go through quality control and are indexed with MeSH or converted to out-of-scope status.
- Citations to articles that are out-of-scope (e.g., covering plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and general chemistry journals, for which only the life sciences articles are indexed with MeSH.
- "Ahead of Print" citations that precede the article's final publication in a MEDLINE indexed journal.
- Citations that precede the date that a journal was selected for MEDLINE indexing (when supplied electronically by the publisher).
- Pre-1966 citations that have not yet been updated with current MeSH and converted to MEDLINE status.
- Citations to some additional life sciences journals that submit full text to PMC® (PubMed Central®) and receive a qualitative review by NLM.
- Citations to author manuscripts of articles published by NIH-funded researchers.
- Citations for the majority of books available on the NCBI Bookshelf (a citation for the book and in some cases each chapter of the book)."

MEDLINE does not include any information about citing or cited references. However, there is information in PubMed about PMC articles cited by other PMC articles.

SCOPUS

Scopus is a database produced by Elsevier Science Publishers. It covers over 18 000 peer-reviewed journal titles from more than 5 000 international publishers and includes 245 million references from the reference lists of the publications covered by

the database. Titles from all geographical regions including non-English titles are considered for inclusion as long as English abstracts can be provided.

Scopus covers several different disciplines such as chemistry, physics, mathematics and engineering, life and health Sciences (including references retrieved from MEDLINE), social sciences, psychology and economics, biological, agricultural and environmental sciences and general sciences. (Scopus Content Overview, 2014)

In Scopus, reference lists are included from 1996, and these are automatically matched to corresponding publication records to obtain the citing/cited information of included publications. Scopus data is sometimes used in bibliometric studies and Elsevier supplies a suite of analyzing tools called SciVal that is based on Scopus data.

GOOGLE SCHOLAR

Google Scholar, which is produced by Google Inc., also contains a kind of citation information. However, the information about citing references is rather unreliable and may contain both duplicates and mismatches.

LOCAL REPOSITORY

A local repository usually does not contain any citation information, but it can contain other information that can be used in a bibliometric study. Depending on the policy on entering publication information, a local repository may well be the most complete record of publications from that particular research unit.

At present (2014), the Karolinska Institutet bibliometric database is directly based on the Web of Science databases (produced by Thomson Reuters) and MEDLINE (a database of the US National Library of Medicine). The system is limited to records available in those databases from 1995 and forward. Articles indexed by PubMed or any other database are at the moment available only if they are also included in the Web of Science databases or MEDLINE.

The Web of Science web interface for Karolinska Institutet also includes proceedings papers. These are not available in the Karolinska Institutet bibliometric system.

SELECTING A UNIT OF ANALYSIS

The starting point in a bibliometric analysis is to select a group of publications. This selection of publications forms the unit of analysis.

The publications may for example be selected on the basis of the authors' organizational affiliation, such as:

- Research group
- Department
- Research centre/Network
- University
- Country

A substantial amount of local data preparation and verification is necessary in order to create a unit of analysis based on a research group, a department, a research centre or a research network. This information is very difficult to locate in bibliographic databases and may in many cases not be present at all. It is even difficult to attribute

publications to a particular university since both organization names and their addresses may be written in many different ways and two different universities occasionally share a common name.

A unit of analysis can also be selected based on the properties of individual articles (instead of authors or author affiliations).

- Individual publications
- Journal
- Subject – often based on subject classification of the journal
- Document type – article, review, note, letter, conference proceeding, etc.
- Publication year

Since statistical methods are used in bibliometric research, the results improve with larger units of analysis. This is partly because isolated phenomena – such as negative citations – are cancelled out by the large amount of articles (Moed, 2005, p. 80). Using bibliometric indicators based on any unit of analysis that contains less than 50 articles (as an individual researcher or group) is not to be recommended (Moed, De Bruin & van Leeuwen, 1995, p. 411).

It is also necessary to take into consideration any possibility of a systematic bias. This could for example be different citing traditions or conventions for including and ranking authors that vary significantly between different research areas (Moed, 2005, p. 223).

The Karolinska bibliometric system includes a "verification tool" where all our researchers are requested to go in and verify their own publications and supply information to where they were active when the publication was written. Since data from the bibliometric system is used for resource allocation to the departments at Karolinska Institutet and the Stockholm County Council most authors log in and verify at least once a year. There is also a module called "analysis toolkit" where researchers can see analyses results for their own publications. This increases analysis transparency.

Karolinska Institutet is very restrictive when it comes to doing analyses on individual researchers or smaller groups. The general guidelines that we follow when doing an analysis on small aggregates are available at http://kib.ki.se.

CHOOSING TIME AND CITATION WINDOWS

In any bibliometric study, it is necessary to decide what time intervals should be used as the basis for the data collection process.

First, the *time window* has to be decided, i.e. the time over which you want to study the unit's publication performance, i.e. the years in which the studied articles were published.

Second, you have to decide between which years the subsequent articles that are going to deliver citations to the analyzed articles have to be published to be included in the citation count. This is called the *citation window*.

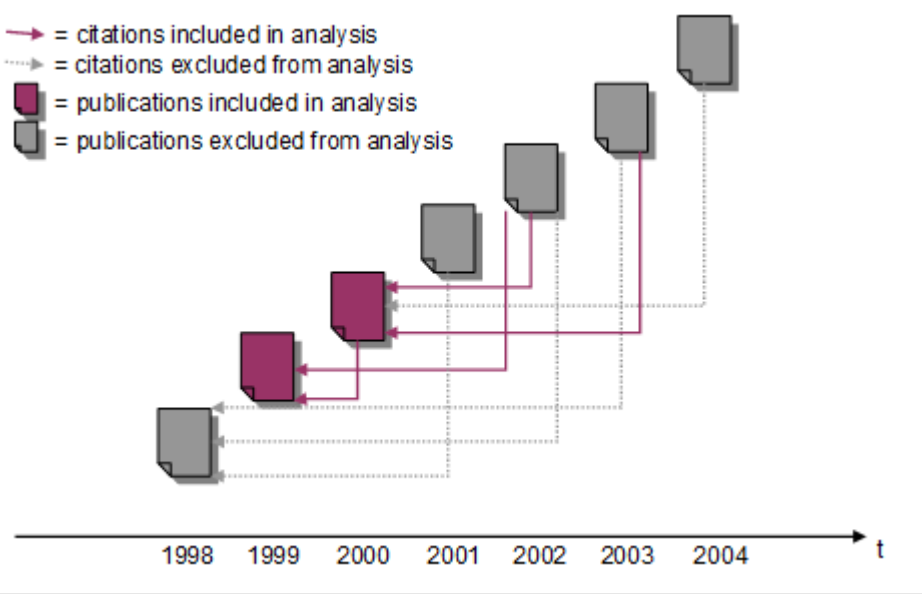

Figure 12. The time window (publication years for cited publications) is 1999–2000 and the citation window (publication years for referring (citing) publications) is 2000–2003.

TIME WINDOW

A time window of 8–10 years (two PhD student generations) can be considered sufficient to analyze the publication activities of one individual research group (Moed 2002, p. 38). When analyzing larger units, the importance of the length of the time window diminishes, but it should of course be the same for all analyzed units.

CITATION WINDOW

There are a few different types of citation windows, and they will affect the analyses results differently.

A citation window can either be fixed or variable. A fixed citation window is set to a specific time period, for instance three years to include the citation peak of most disciplines. A variable citation window uses differently sized citation windows for publications from different years, the citations to all publications are for example counted up to and including 2004, regardless of publication year.

Fixed citation windows are usually "overlapping", that is, they depend on the year of publication. If an analysis uses fixed, overlapping citation windows of for example three years, citations to a publication from 1999 would be counted 1999–2002 and citations to a publication from 2000, 2000–2003. This gives a somewhat fairer image if you wish to compare citation counts of publications from different years. (An alternative could be normalized citation counts, see below.)

A fixed non-overlapping citation window would have the same "citation years" for all publication years included in the specified time window, for instance, citations from 2000–2003 will be counted both for publications from 1999 and 2000. This is of course a disadvantage for the more recent articles when it comes to citation counts.

A long fixed citation window will exclude newer publications since these will not have had time to reach the end of their citation window. A short fixed citation window may cause several normalization groups to have too few citations to be suitable for analysis.

When using an indicator that is normalized with regard to publication year, the citation window will usually be variable. If you plan to use basic indicators like the raw numbers of citations, you have to decide if you wish the citation windows to be fixed and overlapping/non overlapping or variable.

At Karolinska Institutet we mainly use an open citation window. Some specific analyses however require fixed citations windows, such as finding the share of articles that are still uncited three years after publication.

Since we usually analyze relatively large units, it is possible for us to use a time window of five years and still have a sufficient number of publications. The most recent five publication years is therefore our most frequently used time window.

2.2 TECHNICAL ISSUES WITH DATA

MISSPELLINGS

Any type of misspelling in the data, be it of author names, addresses, journal titles or other information, will lead to incorrect numbers of citations and publications. This type of error is fairly common and may be both due to misspellings by the original authors or by mistakes made by the database producer. In large amounts of data however, the effect of these errors is often negligible.

In the Karolinska bibliometric system we show authors, titles etc as they are written in the original databases MEDLINE and/or Web of Science. Corrections are not directly entered into our local database, but we frequently report errors to the database hosts and have good experiences in getting them corrected in their data deliveries within a reasonable time.

UNKNOWN ADDRESSES

As much as 2.4% of the articles, letters, notes and reviews in the Science citation index 1993–2003 and 14% in the Social Science Citation Index (excl. SCI) lack information about authors and author addresses. (Moed, 2005, p. 186) The percentage is higher if other document types are included. These references will automatically be excluded from many types of analyses.

CONNECTIONS BETWEEN REFERENCES AND THE CORRESPONDING ARTICLE

Automatic matching of reference lists to corresponding articles always fails to identify some of the connections. Moed (2002, p. 731) says:

> "when data are derived from 'simple' or 'standard' citation-matching procedures, citation statistics at the level of individuals, research groups, journals and countries are strongly affected by sloppy referencing, editorial characteristics of scientific journals, referencing conventions in scholarly subfields, language problems, author-identification problems, unfamiliarity with foreign author names and ISI data-capturing conventions."

Moed found the overall number of discrepant cited references in the Web of Science to be about 7% but it may be much higher in specific situations.

2.3 ASPECTS OF INTERPRETATION

Several aspects of an essentially non technical nature have to be considered before it is possible to perform a bibliometric analysis, or indeed, interpret the results of one.

IDENTIFYING AN ORGANIZATION AS THE UNIT OF ANALYSIS

It is not always easy to define an organization through the author addresses. The addresses are written by the authors themselves, and it is not uncommon for them to write either the department name or the university name instead of both. In some cases only the main address of a great consortium of writers is written on the paper, and the addresses of the individual authors are left out. Many organizations also have several different units with separate addresses, such as a university, the attached medical school and the university hospital (van Raan, 2000). De Bruin and Moed (1990) have described some of their work on unifying addresses.

In fact, manual identification seems to be the only way to get reliable data on the publishing organizations. Most analysts produce advanced search strings to identify publications from specific organizations with a reasonable amount of work.

You also need to decide if the publications of a university are to be defined as the publications of the people attached to the organization or the publications that were published while someone was working within the organization. This decision depends on whether you wish to assess what the unit has already achieved (an "organization based" study) or what it has the potential to achieve (an "author based" study).

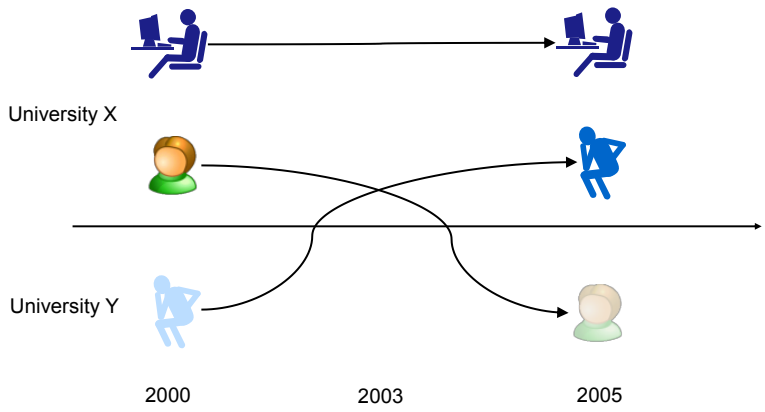

Figure 13. An organization based study only includes articles published during the time an author was working at the analyzed unit. Articles published before moving to a unit or after moving from a unit are disregarded.

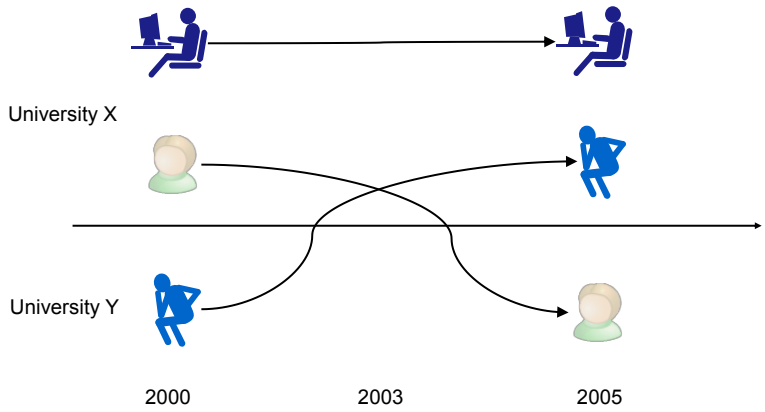

Figure 14. An author based study includes all articles published by authors still working at the analyzed unit, including articles published before moving to the unit. Articles published by authors that have moved from the unit are disregarded, even if they were produced while working at the unit.

The organization based study will depend on correct identification of the organizations in the address field, the author based study on the identifications of the names in the author field.

Address curation is both cumbersome and time consuming. We do not unify all addresses in our database, so the level of data curation for our analyses differs. For about fifty Swedish and international organizations we have developed detailed profiles that are used regularly, but mostly addresses are unified for each specific analysis.

CATEGORIZING PUBLICATIONS INTO RESEARCH FIELDS

Many bibliometric analyses and indicators use some kind of categorization of the analyzed publications into research fields. This is usually based on the subject headings of the included articles or journals supplied by the database producer. At present, most bibliometric analyses do not have any information about the subject of an individual article, only the subjects of the journals that the articles are published in, and this is used as a proxy for "article-subject". Often, the Thomson Reuters classification scheme for journals is used, since this is readily available and covers many different disciplines.

For indicator normalization purposes, at Karolinska Institutet we generally use the Thomson Reuters journal categories. For other types of analysis however, the approach is more varied. Since Medical Subject Headings (Medical Subject Headings 2014) are available for most of our publications, we frequently use search profiles with MeSH to identify publications in specific areas. We also generally use MeSH in author-subject copublication maps.

SELF CITATIONS

The expression "self citations" can be used both for an author or a unit citing their own papers and for publications in a journal citing publications in the same journal. The first of these is the more common usage, and the one used in this handbook (unless specified otherwise).

Studies have shown that self citations do not significantly influence analysis results when you study a sufficiently large number of publications (Glänzel, 2003, p. 57). This is probably because most researchers refer to their own work in equal quantity as a natural part of scientific communication. On group level however, small differences in citation counts may indeed influence indicators. The small number of publications increases the possibility of extreme indicator values.[1]

You can address the aspect of self citation by:

- trying to exclude self citations when calculating indicator values;
- noting them so that the interpretation of the indicators can be affected by the amount of self citations;
- assuming that they are evenly distributed and hence ignoring their effect when calculating the indicators.

It is very difficult to remove self citations when calculating indicators and it requires data from a comprehensive citation database such as the Thomson Reuters citation indices and usually also verification by the analyzed unit.

At Karolinska Institutet the present praxis is not to remove self citations, except for quality assurance purposes. Most of our analyses are made on sufficiently large aggregates that the self citation rate can be expected to "equal out".

FRACTIONALIZATION AND WEIGHTING

Not all authors can be considered to have played an equally large part in producing a publication and some publication types may be considered to be of more importance than others in a particular research area. Also, not all citations are always considered to be equal. There are therefore different ways of compensating for this and similar factors when calculating the number of publications or number of citations attributed to a certain author or unit.

Fractionalization means that only a part of a publication and/or a citation is attributed to the unit of analysis according to some mathematical frequency principle.

Weighting means that some types of publications and/or citations are considered to be of more importance than others and therefore given a higher "weight". Implementation of weighting introduces a subjective element into the analysis since the factor by which the publication/citation is to be weighted has to be decided by the analyst in collaboration with experts in the research fields that are to be evaluated.

The impact of fractionalization and weighting on indicators may be considerable. One should always look at what kind of behaviour will benefit or be at a disadvantage with the different ways of counting and take this into consideration when studying the indicators. Fractionalization and weighting according to the number of institutions in the affiliation list may for example discourage inter-institutional cooperation, something that may not at all be a desired effect.

[1] The aspect of self citations has been left out in the indicator descriptions in the appendix *Bibliometric indicators – definitions and usage at Karolinska Institutet.*

FRACTIONALIZING AND WEIGHTING THE NUMBER OF PUBLICATIONS

If you do not fractionalize or weight the number of publications in any way, each publication is attributed to each of the authors and each publication will be considered to be of equal value. An article with five authors will for example be attributed once to each of the five authors and thus counted as "published" five times. The sum of all articles by all individual authors will thus be larger than the total number of individual articles. This way of counting is often called *full* or *integer* counting.

There are several different methods for fractionalizing and weighting publications:

- *Fractionalizing by authors or other producing units.* Fractionalization by giving each of the authors an equal part of one publication, if for instance an article has five authors, each author can count 0.2 as his/her article. The same principle can be used when attributing publications to institutions instead of authors.
- *Weighting by authors or other producing units.* Weighting according to place in the author list. In some disciplines, for example medicine, the author that has done most of the work is often first in the author list and the most senior researcher is last. You could for example give the first author 50% of the publication, the last author 20%, and divide the remaining 30% between the rest of the authors. However, this differs not only between disciplines but also between countries, departments and journals which makes comparison with other units rather uncertain.
- *Weighting by document types.* In some areas, especially those where other types of publications than journal articles are important, you can also consider giving more weight to some publication types, for instance let books count twice as much as journal articles.
- *Fractionalizing by field.* In you are doing an analysis that produces separate indicators for each included scientific field, the number of publications can be fractionalized between different fields. A publication categorized as belonging to both oncology and haematology will for example contribute half a publication to each field.
- *Weighting by field.* If a publication can be considered to belong more to one field than to the others that it has been classified with, this field can receive a higher share of the publication than the other fields. This is very difficult to implement with a categorization made with the Thomson Reuters journal subject categories but may be possible when using for example MeSH terms for article classification since these include a division of terms into major, ordinary and subheadings.

FRACTIONALIZING AND WEIGHTING THE CITATION COUNT

- *Fractionalizing and weighting by authors (or other producing units), document types and fields.* The citations to an article can be fractionalized and weighted with the same methods and considerations as the number of publications.
- *Fractionalization according to length of reference list.* When an article refers to another article, it can be considered that the value of the referral stands in inverted proportion to the length of the reference list. In other words, a

referral from an article with a long reference list could be considered to be of less value than a referral from an article with a short reference list. This aspect can be imposed by fractionalizing the value of citations by dividing the citation value with the number of items in the reference list. For instance, a citation from an article with 20 references would give the referred item a citation value of 1/20.

- *Weighting according to citation source.* Citations to an article may come from many different sources, and the sources cannot always be considered to be of equal importance. It may be desirable that citations from articles in journals with high impact indicators should be weighted higher than those from articles in journals with low impact indicators.
  A special kind of citation weighting according to source is the Google PageRank ranking algorithm. In short, this algorithm gives higher weight to citations from publications that have themselves a high citation count. There have been some successful experiments that try this algorithm for weighting of the importance of citations to journal articles (Chen, Xie, Maslov & Redner, 2006).

The standard at Karolinska Institutet is not to use fractionalization or weighting of either publications or citations. When doing internal analysis it is very important not to introduce unwanted incentives for the researchers, and both methods have the potential to create such incentives. Fractionalization can for example be an incentive for reducing cooperation.

Although fractionalization is a standard method, for example to counteract the influence of high citation counts because of international collaboration, we believe that fractionalization may give an unwarranted impression of being more "fair". It implies that more authors mean less effort per author, but at the same time there is no way of knowing from the lists of authors and affiliations which authors or organizations have contributed the most. In summary, fractionalization makes the method more difficult for our users to comprehend, without necessarily introducing "fairness".

We find that one major drawback of not using fractionalization is that publications often risk being counted "twice". This means that aggregating publications from several organizations would result in an inflated publication count, which always has to be noted in analyses results. Additionally, since fractionalization is not done for publications classified as belonging to more than one field, the world average for normalized indicators such as the field normalized citation score is inflated to a value somewhat above one.

Another major drawback is that the actual indicator values differ in size between our analyses and analyses made by other bibliometric analysts. Our view is that a bibliometric indicator cannot be seen as an exact value of quality or impact. The indicator value itself is only relevant when compared to other, similar units or to study the development of one unit over time. Also, the aim of most of our analyses is to study the exchange of intellectual property, rather than to quantify research effort.

However, we find that the actual analyses conclusions are seldom affected when fractionalized and unfractionalized results are compared, and that the advantage of not producing unwanted incentives outweighs the drawbacks.

LANGUAGE

An analysis of Web of Science data shows that non-English publications on average get cited much less than publications in the English language. The indicator values of non-English authors improve when you exclude their non-English publications (van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003). Excluding non-English publications may thus be an option when using bibliometrics to assess the research performance of non-English researchers.

CITING TRADITIONS

The citing traditions within different fields affect the number of citations given to different articles. Different fields also vary in how quickly a paper will be cited, how long the citation rate will take to peak and how long the paper will continue being cited. Eugene Garfield describes several factors that contribute to a field's "citation potential". One factor is the length of the reference lists, which for example is twice as long in biochemistry as in mathematics. Another factor is the size of the fields "core literature" (Garfield, 1979, p. 248). Many disciplines connected to medicine and life sciences have a high citation potential.

It is possible to get an overview of the citing traditions in a particular field by looking at the reference lists for a selection of articles within that field. The length of the reference lists, the subjects of the cited publications and their publication years to some extent give an insight in the citation speed and behaviour of that particular field.

SKEWED DISTRIBUTION OF CITATIONS

The distribution of citations to publications is by no means linear, even for the articles of one single author. We have already mentioned Henk Moed's theory about "brick" and "flag" papers and a closer description of this theory is available in his book "Citation analysis in research evaluation" (Moed, 2005, pp. 216–218). For many statistical calculations however, you need a curve that is at least approximately linear. This can partly be achieved by using a logarithmic scale for one, or both, of the diagram axes. For more information, see Seglen (1992). The skewed distribution is something that needs to be taken into consideration when interpreting all bibliometric indicators.

COVERAGE

The quality of a bibliometric analysis improves with increasing coverage of publications in the area you wish to study. Moed (2005) has studied the coverage of the Thomson Reuters citation indices and supplies information on coverage indicators and how to assess a unit with insufficient coverage in the Thomson Reuters citation indices.

The estimated coverage of Karolinska Institutet peer reviewed articles in the Karolinska bibliometric database is over 90%.

IS A BIBLIOMETRIC ANALYSIS ADVISABLE FOR A PARTICULAR UNIT?

Before a bibliometric assessment is undertaken, there are some things to consider:

- You have to estimate how large a share of the unit's publications that can be found in the databases available to you, and if this share is sufficiently large to make bibliometric analysis a reliable option. Is for example the international English serial literature the unit's main mode of communicating its findings? If not, the analysis cannot be made using an already existing citation database, and indicators based on citations can therefore not be included.
- Is it possible to, manually or automatically, identify the papers that belong to this particular unit in the database you chose as your data source? The smaller the number of publications for a unit, the larger the consequences if one or a few of these publications are missed.
- Is the total number of publications by this unit sufficiently large to produce reliable indicators? A small number of publications can produce very extreme indicator values.
- Who will receive the results of the analysis? Interpretation of the indicators can be difficult and continuous discussions with the person or persons requesting the analysis are necessary in order to agree on the best indicators in the individual case and what these may show, and indeed, not show, with regard to the analyzed units.

CHOOSING BIBLIOMETRIC INDICATORS

Many bibliometric researchers stress the importance of not considering the results from any bibliometric analysis to be "truths". Bibliometric methods contain so many simplifications that they only supply a very limited picture of the research they are trying to describe.

No bibliometric indicator should be put to isolated use. Several indicators should always be combined to achieve a more comprehensive picture of the scientific production of a unit (van Leeuwen et al., 2003). A field normalized citation mean indicator should for example be accompanied by information about whether the mean value of citations to the unit's publications is due to a few very highly cited articles or a majority of publications cited a bit above average, and by a quantity indicator to show how many publications that are included in the analysis.

It is important to see bibliometric indicators as one of several tools to be used by competent reviewers with specific knowledge about the research areas included in the analysis. This is for example evident when publications containing very new or

unconventional research results are included in an assessment. These will not yet have been cited, which means that any assessment based solely on bibliometric indicators will not discover the possible potential of the research groups in question.

Within Karolinska Institutet and the Stockholm County Council (SLL) there are recommendations for if, and how, bibliometric methods should be used to analyze individual researchers or small groups.

It is unusual for one author to achieve a publication quantity sufficient for the results to be reliable and stable. It is also important that analyses methods do not create undesirable incentives for publication and verification behaviour, and one expressed intention with bibliometric analyses within KI/SLL is that verification of a publication should never be counterproductive for an individual researcher.

However, with good knowledge of the limitations existing at the level of the individual, certain bibliometric results, preferably non-numeric, are sometimes used to supplement visual inspection of publication lists.

The full KI/SLL recommendations for bibliometric indicators suitable for individuals or smaller groups are available at http://kib.ki.se/.

2.4 BIBLIOMETRIC INDICATORS

This chapter contains some examples of the most commonly used and/or important bibliometric indicators. For a more complete list, see the document *Bibliometric indicators – definitions and usage at Karolinska Institutet*.

BASIC BIBLIOMETRIC INDICATORS

Basic bibliometric indicators are characterized by being mainly raw publication and citation counts, where there is no or little compensation for the size of the analyzed unit, nor are the characteristics of citation patterns regarded.

NUMBER OF PUBLICATIONS AND CITATIONS

Two very basic bibliometric indicators are the number of publications and citations during a specific time period. These two indicators do not compensate for the size of the publishing unit or the document type of the publications. However, they can be useful to someone with knowledge of the research area under study, especially if the indicators are used to compare similar research units or as a complement to other bibliometric indicators.

Since most publications are written by more than one author, it is not always straightforward how you should count the number of publications for each author. Different kinds of weighting schemes for publications and citations have been described earlier in this handbook.

NUMBER OF PUBLICATIONS AND CITATIONS PER RESEARCHER

Publication and citation counts in relation to the number of active researchers or employees at the studied unit are two somewhat more refined indicators of scientific production and impact. It can however be surprisingly difficult to find out the number of active researchers at one particular unit.

CITATIONS PER PUBLICATION

The average number of citations that articles published by an analyzed unit during the analyzed time span has received.

It gives an indication of the average scientific impact of a unit's published articles, but it does not take into account that older articles usually are more cited and that citation rates vary between document types and subject areas.

H-INDEX

The h-index is the maximum number of publications (h), attributed to an analyzed unit during the analyzed time span, that have at least h citations. In its original form it is calculated for individual researchers and based on their entire publication output.

It is very easy to calculate in the Thomson Reuters Web of Science, but it has the same disadvantage as the other basic indicators – it compares documents of different types, published in different years, in totally different subjects, with each other.

The h-index gives positive bias to senior researchers with older articles, since these have had more time to be cited. Another criticism on the h-index is that it disfavours scientists with a short career, since the h-index never can be larger than the number of

published articles, no matter how important and well-cited the articles are. In short: a researcher with a few extremely highly cited articles will still have a low h-index, and a researcher with a high h-index could have stopped publishing his best material several years ago.

THOMSON REUTERS IMPACT FACTOR

The Thomson Reuters Impact Factor was designed by Eugene Garfield around 1960 as a means to measure the scientific impact of a specific journal. It gives an average value on how many times an article in the journal has been cited. It is defined as the average number of citations given in a specific year to documents published in that journal in the two preceding years, divided by the number of documents published in that journal in those two years.

The Impact Factor is used by Thomson Reuters to localize the most important scientific journals in each research area and is in this respect also used as a library collections management tool.

The highest impact factors can exceed 100, and journals like New England Journal of Medicine, Nature and Science have Impact Factors between 30 and 50. Many journals have an Impact Factor below 1.

The fact that the Thomson Reuters Impact Factor is based on citations only 1–2 years old can be considered a compromise between the need of getting a quick appraisal of new journals and letting the publications reach their citation maximum (the year when the publication receives most of its citations). Most articles reach their citation maximum 3–5 years after publication, so that would for many research areas be a preferable citation window.

The citation patterns vary so much between different research areas that the Thomson Reuters Impact Factor should not be used to compare the scientific impact of journals in different subjects.

Since the Thomson Reuters Impact Factor is relatively easy to find and understand it has become very popular and is often used to assess the quality of articles, researchers, departments and universities by studying the journals they publish in. This is unadvisable and not what the Impact Factor was intended for. Because of skewed distributions, the impact of a journal is also not sufficient information to reliably predict the number of citations to an individual article.

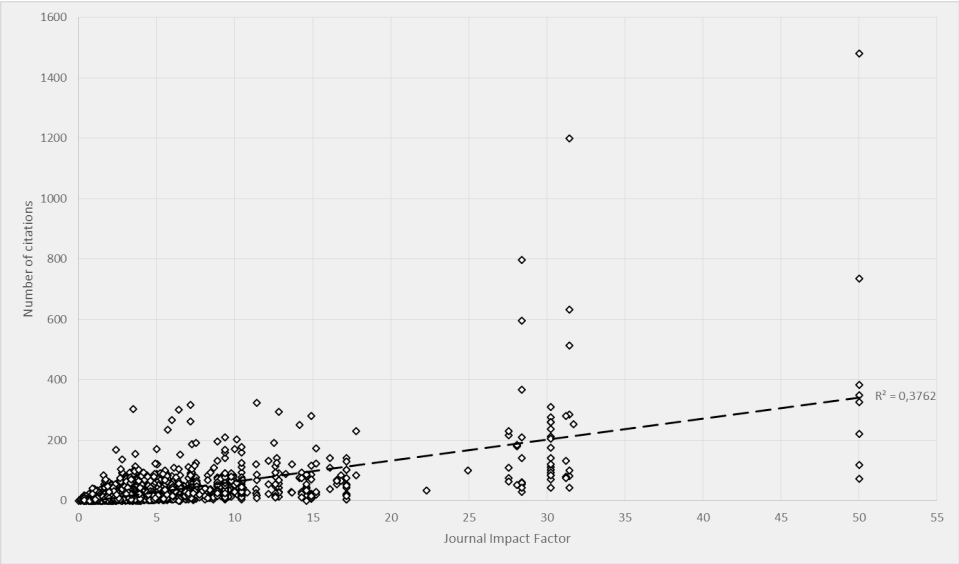

Figure 15. Impact Factors of journals and cumulative number of citations in 2014 based on articles produced by Karolinska Institutet in 2008. Each marker represents an individual article and the dashed line represent a linear regression of number of citations on Impact Factor.

However, if you wish to study very recently published articles that have not yet been cited, a journal impact indicator may be the only possible performance indicator. The indicator is then based on the assumption that the refereeing process is more rigorous in high-impact journals, which means that only high-quality research will be accepted. For any type of analysis that relates to journal impact, it is advisable to supplement the Thomson Reuters Impact Factor with one of the more advanced normalized journal impact indicators described below.

NUMBER OF PUBLICATIONS IN HIGH-IMPACT JOURNALS

Publications in high-impact journals are often considered to be of high quality. It is not uncommon for researchers to be asked to supply information about the average value of the Impact Factors of the journals they have published in, when they apply for a grant or a new position.

Sometimes a research unit or a university also displays how many publications they have in journals with very high Thomson Reuters Impact Factors, the 20 or 40 most highly ranked, as a sign of quality research produced by authors at that unit. This measurement is usually not compensated for the size of the analyzed unit, which means that larger institutions get higher values. The research areas of the analyzed unit will also affect this figure, so sometimes different Impact Factor lists are used for different research areas.

However, as mentioned earlier, the Impact Factor of a journal cannot predict the number of citations that any individual publication will receive.

ADVANCED BIBLIOMETRIC INDICATORS

Advanced bibliometric indicators have three important aspects "built-in".

- Publication year – citations accumulate with age which means that older articles are more highly cited.
- Document type – the number of citations to different document types varies significantly. Review articles, for example, generally receive more citations than regular articles.
- Research area – the citation patterns are different in different research areas

The advanced indicators always include a normalization process, i.e. a comparison of publication citation counts to the citation count average of publications of the same document type, published the same year, in the same subject.

FIELD NORMALIZED CITATION SCORE

The field normalized citation score compares the number of citations to the publications of an analyzed unit to the number of citations to international publications from the same year, in the same research field and of the same document type.

The normalized citation score is usually written as a decimal number that shows the relation to the normalized world average 1, which means that 0.9 shows that the analyzed publications are cited 10% less than the world average and 1.2 that they are cited 20% more.

The field normalized citation score may be calculated in two different ways. The initial field normalized citation score invented by Centre for Science and Technology Studies in Leiden (CWTS), called the "crown indicator", aggregates the unit's publications as a whole, calculates a mean citation value, and then divides that with the average of the field citation scores for the fields, years and document types the aggregated articles belong to.

At Karolinska Institutet, an alternative way to calculate was developed (Lundberg, 2007). It is called the "Item oriented field normalized citation score average", and it is calculated by normalizing each individual publication's citation rate against an average citation rate for articles in the same subject area, the same type and of the same age, and finally the average of all the normalized citation values are calculated.

In 2013, CWTS changed their method for calculating the crown indicator to the item oriented approach (Waltman et al., 2011) and the crown indicator used by CWTS and the field normalized citation score used by Karolinska Institutet is now very similar. The resulting indicator values however are not entirely comparable, since CWTS use fractionalization as a standard whereas Karolinska Institutet (for reasons stated above) do not.

Read more about the variants of this indicator in *Bibliometric indicators – definitions and usage at Karolinska Institutet.*

TOP X%

Top X% shows the number or share of publications attributed to a group of authors that belong to the X% most cited publications in the world from the same year, in the same field and of the same document type.

The Top X% indicator is sometimes written as a decimal number that shows the relation to the normalized world average 1. A value over 1 shows that the analyzed unit has more of its publications among the top X% than the world average, a value below 1 that it has less.

Top X% is often used as a complement to the field normalized citation score. It indicates if a high average citation score is achieved through a few very highly cited articles or a larger number of articles cited above average. It may also identify highly cited articles from a group with a low field normalized citation score whose top publications would otherwise have been unnoticed. Additional (non-bibliometric) information is required to decide if one of the two patterns is a sign of high-quality research.

2.5 PUBLICATION PATTERNS

CO-PUBLICATION

If the publication data used for analysis contains all author addresses (as in data from the Thomson Reuters citation indices) it can be used to find patterns of co-publication between different authors and units. These can be visualized in co-publication tables and maps that show for example the authors, universities and countries that publish together and to what extent.

In most co-publication maps not all authors and connections have been included, since this would make the elements in the map too many for a satisfactory visualization. It is therefore important to know the inclusion criteria for the different elements.

- Number of publications for an author to be included.
- Number of connections for an author to be included.
- Strength of connection for it to show up on the map.

These criteria might be applied one by one or in any combination.

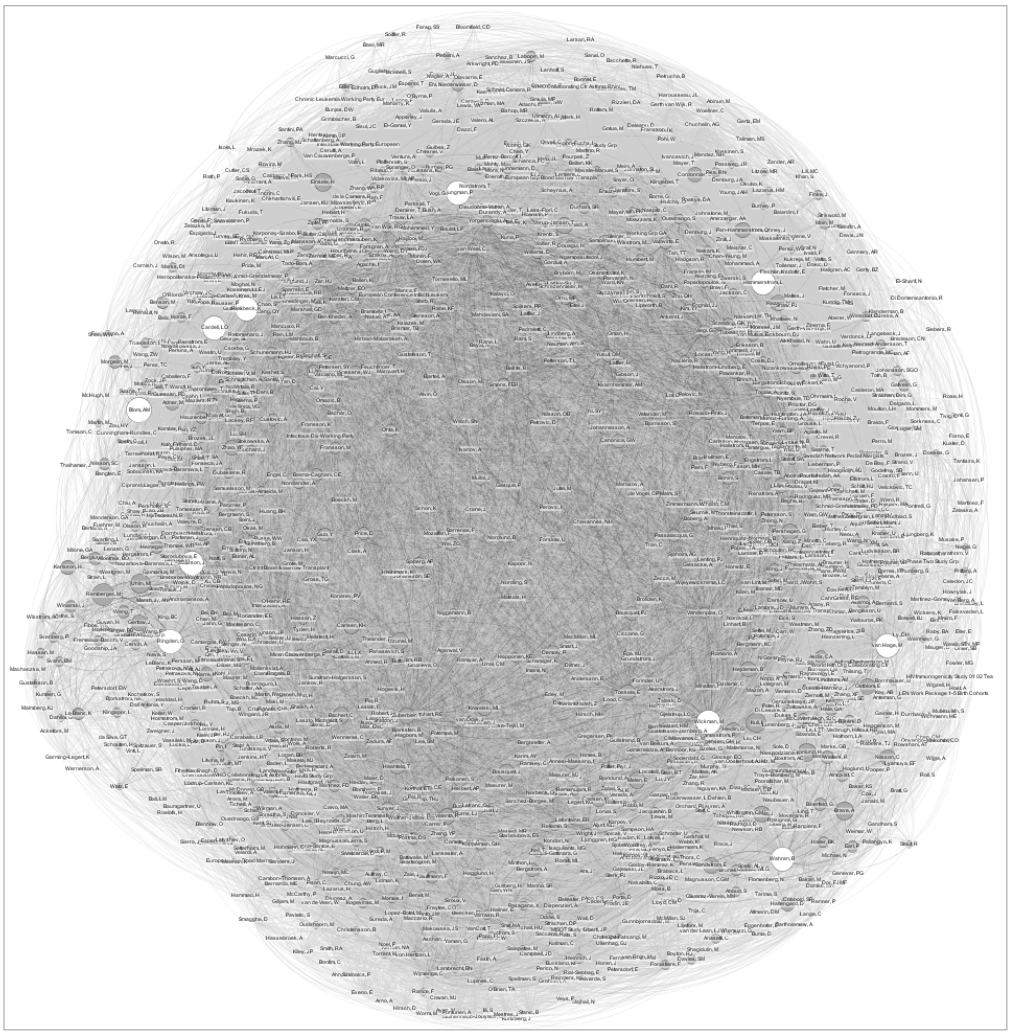

Figure 16. A co-publication map of the ten most prolific Swedish journal article authors (white circles) that have published in immunology journals 2008–2012. All 1 628 authors connected to these are included in the map which makes it impossible to interpret.

If you choose to have a minimum number of publications for each author you will reduce the number of authors by excluding the non-prolific ones. At the same time you also exclude any connections that the authors still present in the map have to any of the excluded authors. You can also choose to disregard connections that are based on less than a selected number of publications.

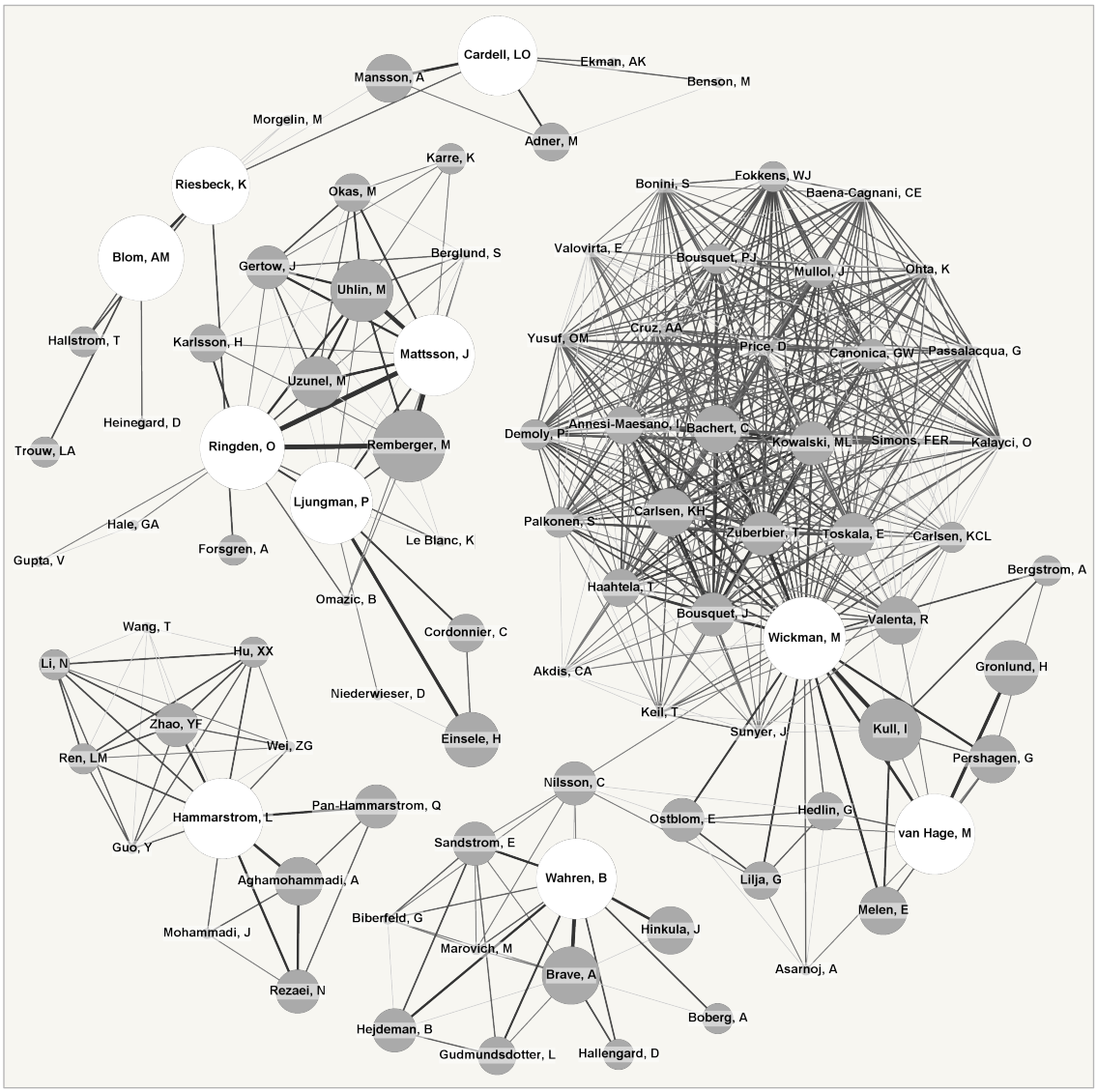

Figure 17. A co-publication map of the ten most prolific Swedish journal article authors (white circles) that have published in immunology journals 2008–2012. Authors need at least five co-publications with one of the ten most prolific, and connections need at least three co-publications to be shown in this map.

If the analysis only shows authors that have many connections to other authors, you will not only exclude non-prolific authors but also authors that only co-publish with one or a few partners.

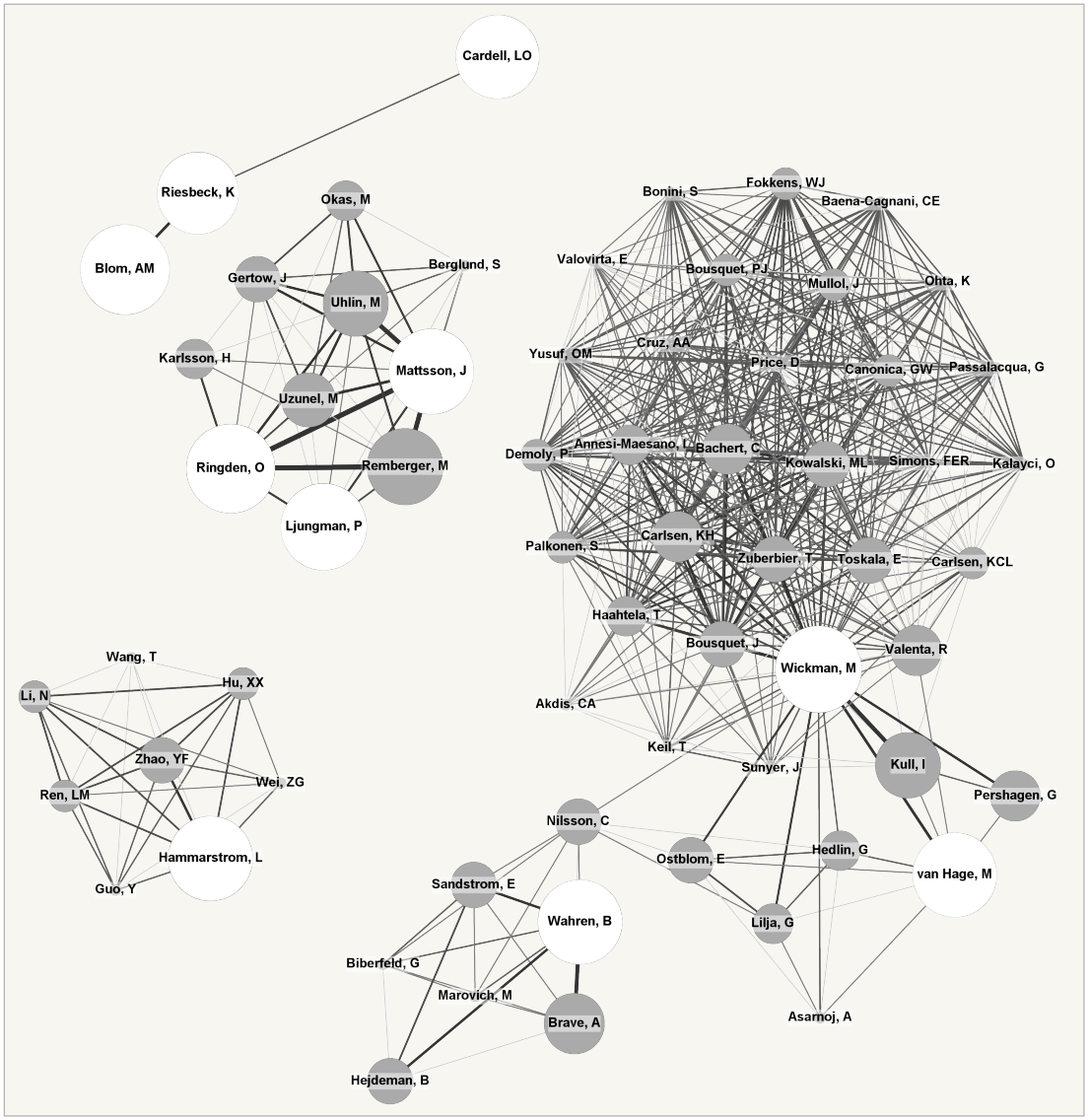

Figure 18. A co-publication map of the ten most prolific Swedish journal article authors (white circles) that have published in immunology journals 2008–2012. Authors need at least five co-publications with one of the ten most prolific, and connections need at least three co-publications to be shown in this map. Authors also need co-publications with at least five other authors to be shown.

SUBJECT INTERRELATEDNESS

If data contains a subject classification of the publications, this can be used to analyze for example:

- Which subjects researchers specify in and how these connect them to each other.
- Research areas that are connected by a certain number of publications.
- If research areas that show up frequently in the publications of a department are organized in one large departmental research network or several smaller ones.

Copublication maps are very frequently requested analyses at Karolinska Institutet and are often used to illustrate how we copublish with other countries and universities.

Due to the frequency of this particular analysis, the bibliometric database has a web interface for collaboration data built especially for Karolinska Institutet employees working with international affairs. This provides maps of Karolinska Institutet copublications with other countries and organizations, including who are the Karolinska Institutet coauthors and what subject areas (represented by MeSH-headings) the copublications are about.

Maps are also made of copublication patterns within and between departments at Karolinska Institutet and the Karolinska University Hospital. We also use network visualizations to illustrate how research areas are connected, or which researchers are active in different areas.

3. REFERENCES

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. Journal of Infometrics, 1(1), 8–15.

De Bruin, R. E., & Moed, H. F. (1990). The Unification of addresses in scientific publications. In L. Egghe & R. Rousseau (Eds.), Informetrics 89/90 : selection of papers submitted for the second International Conference on Bibliometrics, Scientometrics and Informetrics, London, Ontario, Canada, 5–7 July 1989 (pp. 65–78). Amsterdam: Elsevier.

Fact Sheet; MEDLINE, PubMed, and PMC (PubMed Central): How are they different? (2014). Retrieved Mars 19, 2014, from http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html.

Garfield, E. (1979). Citation indexing: its theory and application in science, technology, and humanities. New York: Wiley.

Glänzel, W. (2003). Bibliometrics as a research field: A course on theory and application of bibliometric indicators.

Lundberg, J. (2007). Lifting the crown – citation *z*-score. Journal of Informetrics, 1(2), 145–154.

Lundberg, J., Fransson, A., Brommels, M., Skar, J., & Lundkvist, I. (2005). Is it better or just the same? Article identification strategies impact bibliometric assessments. Scientometrics, 66(1), 183–197.

Medical Subject Headings. (2014). Retrieved Mars 19, 2014, from http://www.nlm.nih.gov/mesh/.

Moed, H. F., De Bruin, R. E. & Van Leeuwen, T. N. (1995). New bibliometric tools for the assessment of National Research Performance – Database description, overview of indicators and first applications. Scientometrics, 33(3), 381–422.

Moed, H. F. (2002). The impact-factors debate: The ISI's uses and limits. Nature, 415(6873), 731–732.

Moed, H. F. (2005). Citation analysis in research evaluation. Dordrecht: Springer.

Oxford English dictionary. (2014). Retrieved Mars 19, 2014, from http://dictionary.oed.com/.

Pritchard, A. (1969). Statistical Bibliography or Bibliometrics. Journal of Documentation, 25(4), 348–349.

Roth, D. L. (2005). The emergence of competitors to the Science Citation Index and the Web of Science. Current Science, 89(9), 1531–1536.

Scopus Content Overview. (2014). Retrieved Mars 19, 2014, from http://www.elsevier.com/online-tools/scopus/content-overview.

Seglen, P. O. (1992). The Skewness of Science. Journal of the American Society for Information Science, 43(9), 628–638.

Science Citation Index Expanded (2014). Retrieved Mars 19, 2014, from http://thomsonreuters.com/science-citation-index-expanded/.

Ulrichsweb.com. (2014). Advanced Search. Retrieved Mars 19, 2014, from http://www.ulrichsweb.com/ulrichsweb/Search/advancedSearch.asp?

United Nations Development Programme Evaluation Office. (2002). Handbook on monitoring and evaluating for results. New York: United Nations. http://web.undp.org/evaluation/documents/HandBook/ME-HandBook.pdf

van Leeuwen T. N. Second generation bibliometric indicators: the improvement of existing and development of new bibliometric indicators for research and journal performance assessment procedures [dissertation]. Leiden: 2004.

van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & van Raan, A. F. J. (2003). Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. Scientometrics, 57(2), 257–280.

van Raan, A. F. J. (2000). The Pandora's Box of Citation Analysis: Measuring Scientific Excellence – the last Evil? In B. Cronin & H. B. Atkins (Eds.), The web of knowledge: a festschrift in honor of Eugene Garfield (pp. 301–320). Medford, N.J.: Information Today.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: an empirical analysis. Scientometrics, 3, 467–481.