

Application of bibliometric analysis in research funding

Thed van Leeuwen
CWTS



Universiteit Leiden
The Netherlands

Contents

- Introduction of bibliometrics, CWTS , and CWTS indicators.
- Methodology of the studies for charities.
- Some results of the studies for charities.
- Some conclusions.

***INTRODUCTION OF CWTS,
BIBLIOMETRICS, & BIBLIOMETRIC
INDICATORS***

Introduction of bibliometrics

- Bibliometrics can be defined as the quantitative analysis of science and technology performance and the cognitive and organizational structure of science and technology.
- Basic for these analyses is the scientific communication between scientists through (mainly) journal publications.
- Key concepts in bibliometrics are *output* and *impact*, as measured through publications and citations.
- Important starting point in bibliometrics: scientists express, through citations in their scientific publications, a certain degree of influence of others on their own work.
- By large scale quantification, citations indicate influence or (inter)national visibility of scientific activity, but should not be interpreted as synonym for 'quality'.

CWTS data system

- CWTS has a full bibliometric license from Thomson Reuters Scientific to conduct evaluation studies using the Web of Science.
- Database covers the period 1981-2008.
- Some characteristics:
 - Over 28.000.000 publications.
 - Over 500.000.000 citation relations between source papers.
 - 48.000.000 authors (incl. variations).
 - 30.000.000 addresses, 90% cleaned up over the last 10 years.
 - Contains reference sets for journal and field citation data.

Some basic indicators are...

- ***P***: number of publications in journals processed for the Web of Science.
- ***C***: number of received citations, excl. self-citations.
- ***CPP***: mean number of citations per publication, excl. self-citations
- ***Pnc***: percentage of the publications not cited (within a certain time-frame !!!)
- **% *SC***: percentage self-citations related to an output set.

Important CWTS indicators are...

- ***CPP/JCSm***: ratio between real, actual impact, and mean journal impact.
- ***CPP/FCSm***: ratio between real, actual impact, and mean field impact.
- ***JCSm/FCSm***: ratio between journal impact, and field impact, indicative for the 'quality' of the journal package in the field

***A MODEL FOR APPLYING
BIBLIOMETRICS BY FUNDING
AGENCIES***

Charities for which CWTS worked ...

- Netherlands Diabetes Foundation
- Dutch Aids Foundation
- Netherlands Heart Foundation
- Dutch Asthma Foundation
- Rheuma Foundation
- MDL-foundation
- Wellcome Trust

Data collection

- Roughly, we can distinguish three standard methods for the collection of a set of publications for an analysis:
 - Based on a list of names of researchers
(verification through a website creates valid dataset)
 - Based on a list of publications of a unit
(the supplied lists form the authorized/verified dataset)
 - Based on the address of an institute or unit
(this approach does not allow lower level insights)
- Combine this with information from a mono-disciplinary database
(e.g., Medline/PubMed, Inspec, etc.)

Combining the collected data

- If we combine data collected from the multidisciplinary WoS with data from a mono-disciplinary database, we create a 'best-of-both-worlds' situation:
 - *Citation data from WoS*
 - *All addresses from WoS*
 - *Disciplinary information from other systems (e.g., key-words, classification terms, etc.)*
- Ideally, one would combine this with information from the funding agency (e.g., on rewarded and non-rewarded grants).

Working of the model

Funding agency

<i>Information on the research field</i>	<i>Information on funded research (known research(ers))</i>		<i>Unknown research(ers)</i>
	<i>Granted</i>	<i>Not granted</i>	<i>Not granted</i>
<i>In the field</i>	<i>Field-related, supported by the agency</i>	<i>Field-related, not supported by the agency</i>	<i>Field-related, not supported by the agency, not known to agency</i>
<i>Outside the field</i>	<i>Non Field- related, Supported by the agency</i>	<i>Non Field- related, not supported by the agency</i>	

SOME RESULTS OF STUDIES FOR FUNDING CHARITIES

Results for Diabetes Foundation

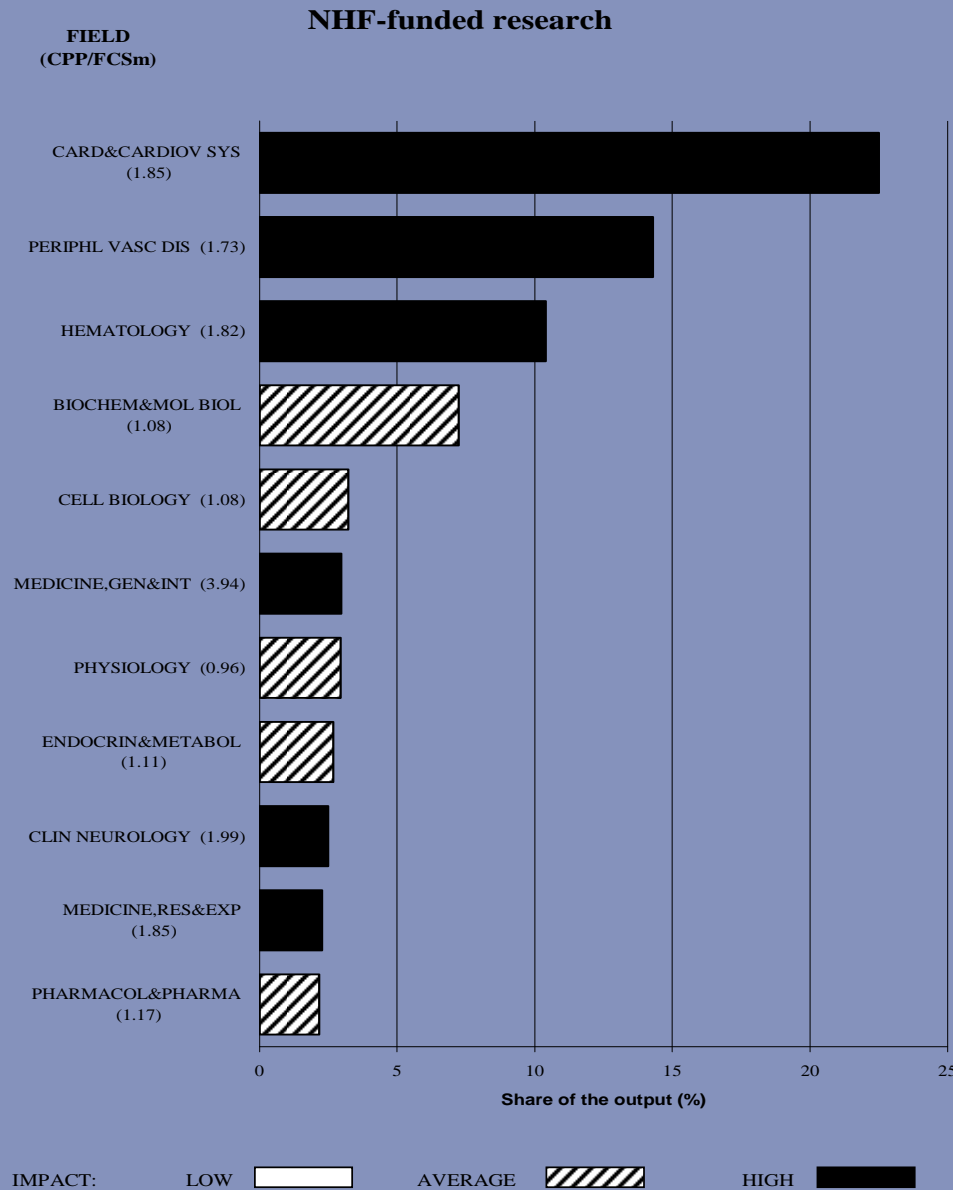
Diabetes field	supported by DFN	not-supported by DFN	not known by DFN
- P:	443	332	150
- C+sc:	3465	2041	2090
- <i>CPP</i> :	6.2	4.7	11.2
- CPP/JCSm:	1.0	0.9	1.3
- CPP/FCSm:	1.1	0.9	1.7
- JCSm/FCSm:	1.1	1.1	1.3
- % not cited:	26%	22%	29%
- % self citations	21%	24%	20%

Results for NHF 1993-2007

	NHF- funded	Resembling NHF- funded
- <i>P</i> :	1.399	87.594
- <i>C+sc</i> :	29.991	1.430.402
- <i>CPP</i> :	16.87	13.26
- <i>CPP/JCSm</i> :	1.07	1.09
- <i>CPP/FCSm</i> :	1.61	1.29
- <i>JCSm/FCSm</i> :	1.51	1.18
- % <i>not cited</i> :	12%	19%
- % self citations	21%	19%

**RESEARCH PROFILE
OUTPUT AND IMPACT PER FIELD
1993 - 2007**

Research Profile



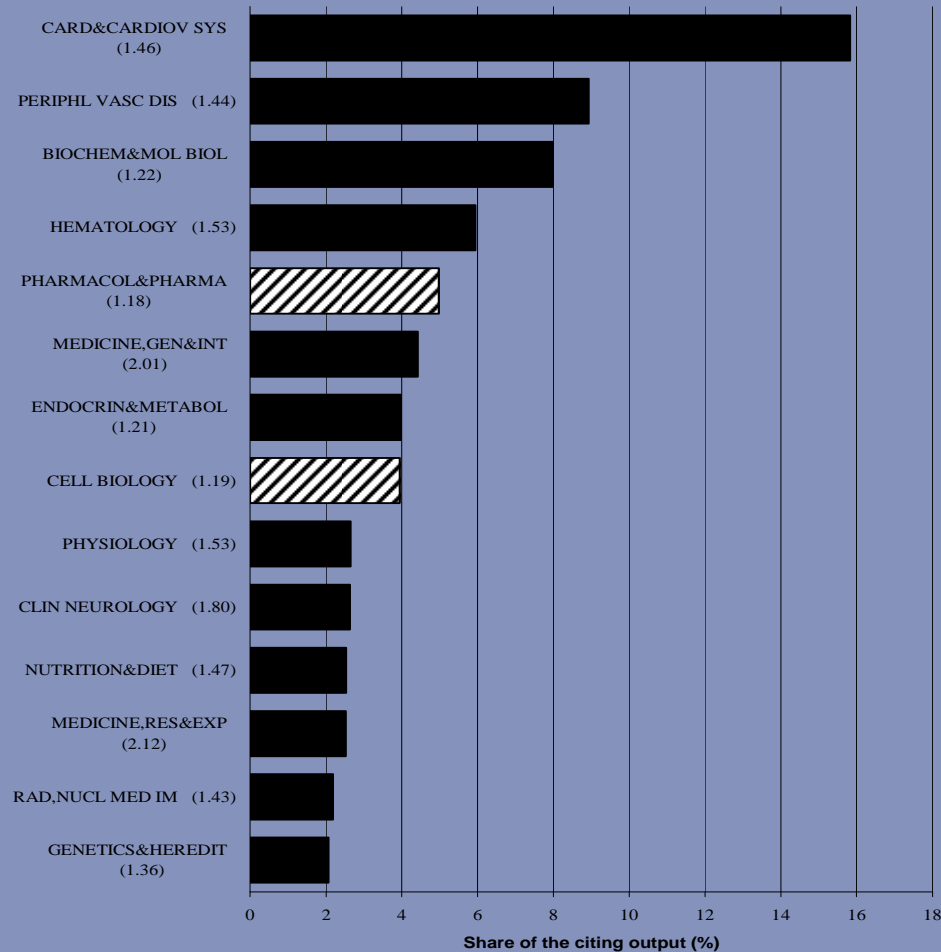
Provides an overview of the fields in which funded research was taking place

Impact measures indicate strong and weak spots in profile

**KNOWLWEDGE USER PROFILE
CITING OUTPUT AND IMPACT PER FIELD
1993 - 2007**

Research citing NHF funded output

**CITING FIELD
(CPP/FCSm)**



IMPACT: LOW AVERAGE HIGH

Knowledge User Profile: Citing fields

Provides an overview of the fields by which the funded research was cited

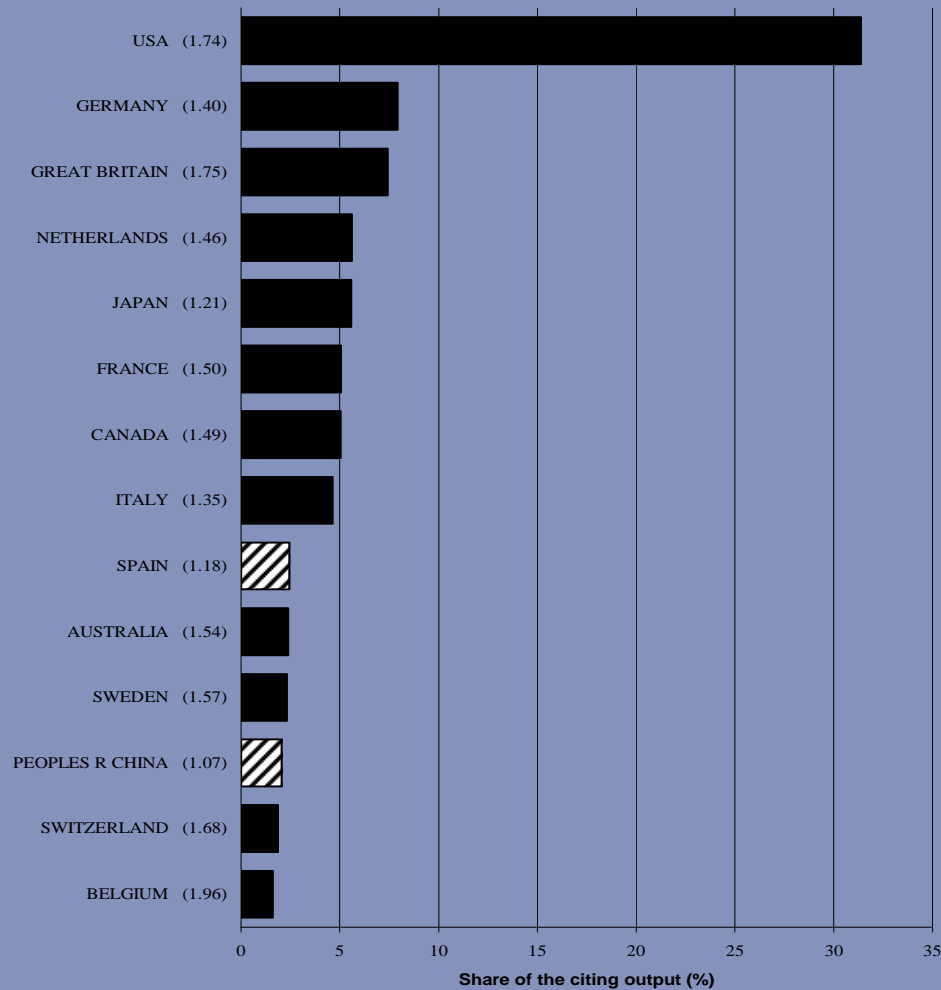
Impact measures indicate the influence on the research front

**KNOWLWEDGE USER PROFILE
CITING OUTPUT AND IMPACT PER COUNTRY
1993 - 2007**

Knowledge User Profile: Citing countries

Research citing NHF funded output

CITING COUNTRY
(CP/FCSm)



Provides an overview of the countries by which the funded research was cited

Impact measures again indicate the influence on the research front

IMPACT: LOW  AVERAGE  HIGH 

Some conclusions ...

- Bibliometrics can be helpful in evaluating the efficiency of the funding strategies of the past.
- As evaluation of research is a matter of governing bodies rather than funding agencies, bibliometrics with light touch peer review will work for charities.
- Mono-disciplinary research has greater changes of getting funded compared to multidisciplinary research.
- Our bibliometric model for funding agencies has proven to be successful in the past.

JOURNAL & FIELD NORMALIZATION



Universiteit Leiden
The Netherlands

Universiteit Leiden. The university to discover.

Calculating the *JCSm* & *FCSm*

	Type	publ. year	Journal	Journal category	# citations until 2008
I	<i>review</i>	2005	CANCER RES	Oncology	17
II	<i>note</i>	2006	J CLIN END	Endocrinology	4
III	<i>article</i>	2006	J CLIN END	Endocrinology	6
IV	<i>article</i>	2007	J CLIN END	Endocrinology	8

Calculating the *JCSm* & *FCSm* 2

	CPP	<i>JCS</i>	<i>FCS</i>
I	17	16.9	23.7
II	4	3.1	3.0
III	6	4.8	4.1
IV	8	4.8	4.1

Calculating the *JCSm* & *FCSm* 3

The mean citation score is determined

as:

$$CPP = \frac{17 + 4 + 6 + 8}{1 + 1 + 1 + 1} = 8.8$$

The mean journal citation score as:

$$JCSm = \frac{(1 \times 16.9) + (1 \times 3.1) + (2 \times 4.8)}{1 + 1 + 2} = 7.4$$

CPP / JCSm

$$(8.8 / 7.4) = 1.19$$

The mean field citation score as:

$$FCSm = \frac{(1 \times 23.7) + (1 \times 3.0) + (2 \times 4.1)}{1 + 1 + 2} = 8.7$$

CPP / FCSm

$$(8.8 / 8.7) = 1.01$$

CITATION WINDOWS & IMPACT MEASUREMENT



Universiteit Leiden
The Netherlands

Universiteit Leiden. The university to discover.

Citation measurement and 'windows'

- **Publication years, fixed citation 'window'.**
Publications of 2001, with three citation years (namely 2001, 2002, en 2003), followed by 2002, with three years, etc.
- **Blocks of publication years with a window decreasing in length.**
Publications of 2001-2004, with citation window of 4 years (2001-2004), 3 years (2002-2004), 2 years (2003-2004), and 1 year (2004), and so forth.

Citation measurement with 'fixed window'

	<i>Citation years</i>							
	2001	2002	2003	2004	2005	2006	2007	2008
2001	<u>2001</u>	2002	2003					
2002		<u>2002</u>	2003	2004				
2003			<u>2003</u>	2004	2005			
2004				<u>2004</u>	2005	2006		
2005					<u>2005</u>	2006	2007	
2006						<u>2006</u>	2007	2008
2007							<u>2007</u>	2008
2008								<u>2008</u>

Citation measurement with 'year blocks'

	<i>Citation years</i>							
	2001	2002	2003	2004	2005	2006	2007	2008
2001	<u>2001</u>	2002	2003	2004				
2002		<u>2002</u>	2003	2004	2005			
2003			<u>2003</u>	2004	2005	2006		
2004				<u>2004</u>	2005	2006	2007	
2005					<u>2005</u>	2006	2007	2008
2006						<u>2006</u>	2007	2008
2007							<u>2007</u>	2008
2008								<u>2008</u>

ISI Impact Factors: calculation and validity



Universiteit Leiden
The Netherlands

ISI Impact Factors

- From 1995 onwards CWTS has analyzed the uses and validity ISI Journal Impact Factor (IF).
- Most important points of criticism were:
 - *Calculated erroneously.*
 - *Not sensitive for the composition of the journal in terms of the document types.*
 - *Not sensitive for the science fields a journal is attached to ...*
 - *Based on too short ‘citation windows’.*

Share 'citations-for-free' for *The Lancet*

	Publications 90+91	Citations 1992
Art	784	2986
Not	144	593
Rev	29	232
Sub-total	957 (a)	7959 (b)
Let	4181	4264
Edi	1313	905
Other	1421	909
Total	7872	14037 (c)

- **ISI Method:**

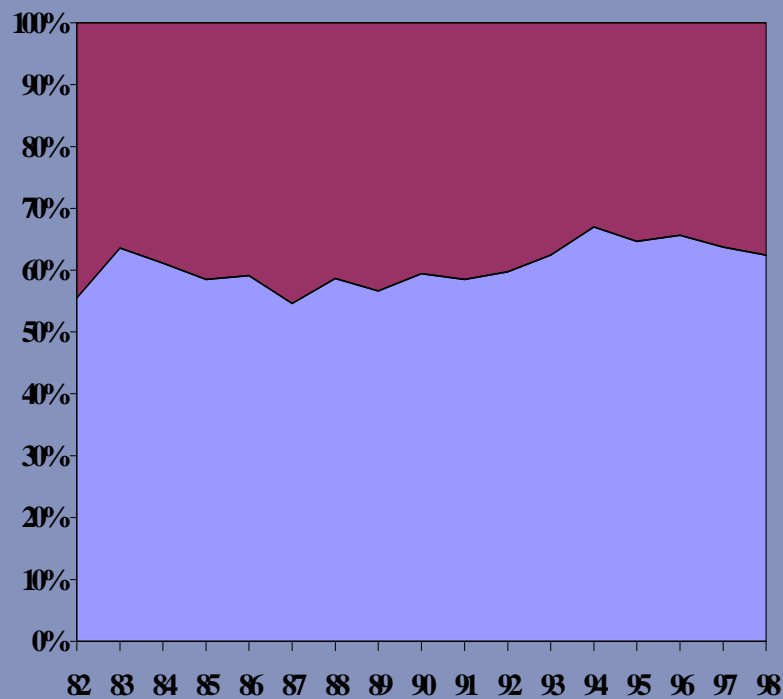
$$\frac{\text{Citations in 2000}}{\text{Citeable documents in '98 and '99}} = \frac{14037 \text{ (c)}}{957 \text{ (a)}} = \mathbf{IF=14.7}$$

- **CWTS Method:**

$$\frac{\text{Citations to Art/Not/Rev in 2000}}{\text{Art/Not/Rev in '98 and '99}} = \frac{7959 \text{ (b)}}{957 \text{ (a)}} = \mathbf{IF=8.3}$$

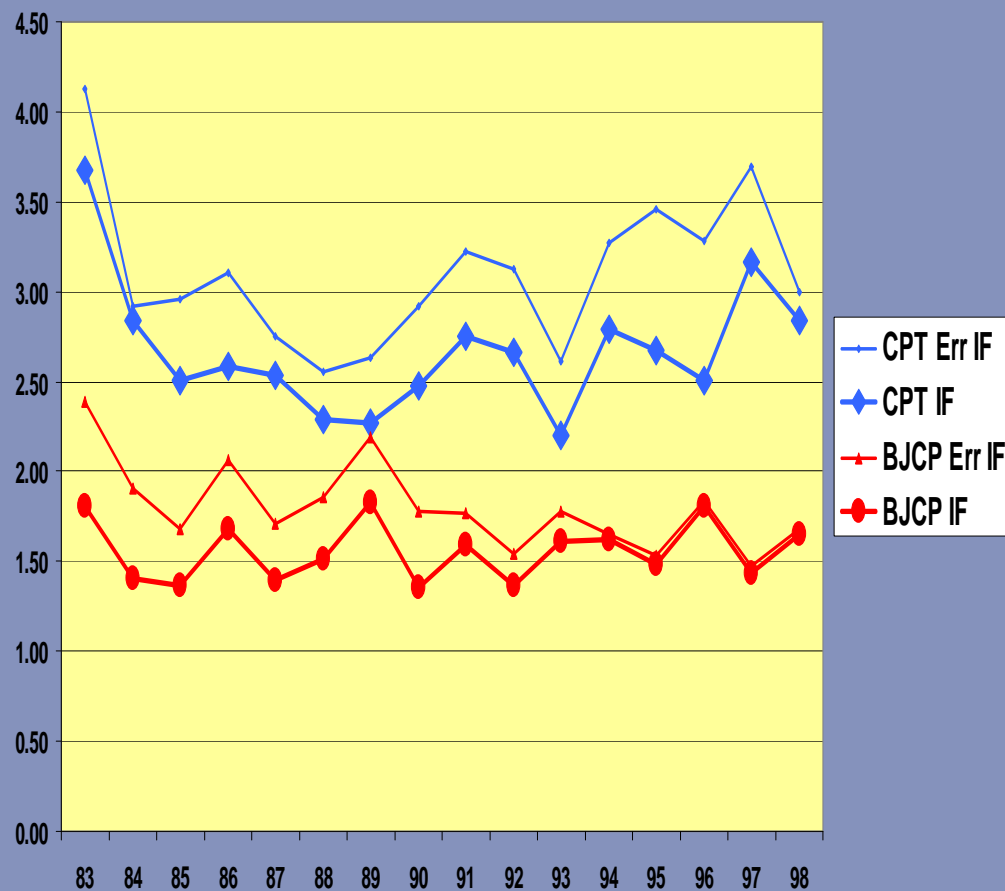
$$\frac{\text{Citations to Art/Let/Not/Rev in 2000}}{\text{Art/Let/Not/Rev in '98 and '99}} = \frac{7959+4264}{957+4181} = \mathbf{IF=2.4}$$

Distribution of citations used for the calculation of the IF value of *The Lancet*



- The red area indicates citations 'for free', while the blue area indicates 'correct citations'
- The IF-score of *The Lancet* is seriously 'overrated' by the scientific 'audience' of the journal.

Impact Factoren voor *Br. J. Clin. Pharm.* en *Clin. Pharm. & Ther.*



- The graph shows the **correct** and *erroneous* impact factors of **BJCP** and **CPT**
- In the case of **CPT**, citations to published **meeting abstracts** are included, while **BJCP** has stopped publishing of **meeting abstracts** !

The H-Index and its limitations

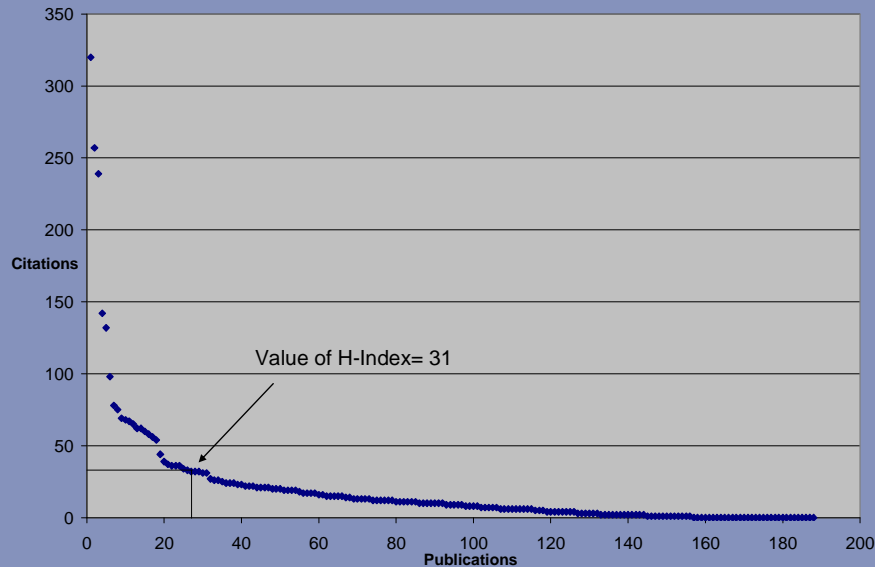


Universiteit Leiden
The Netherlands

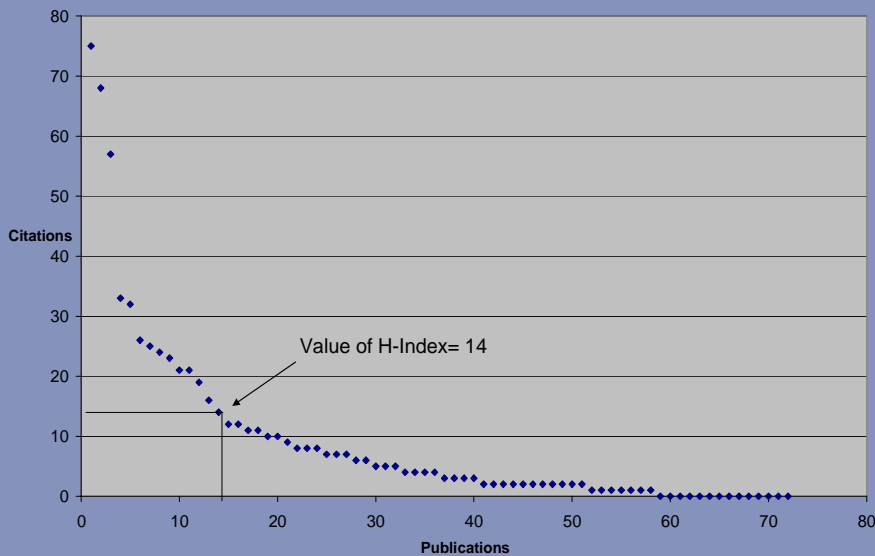
The H-Index, defined as ...

- The H-Index is the score that indicates the position at which a publication in a set, the number of received citations is equal to the ranking position of that publication.
- Idea of an American physicist, J. Hirsch, who published about this index in the Proc. NAS USA.

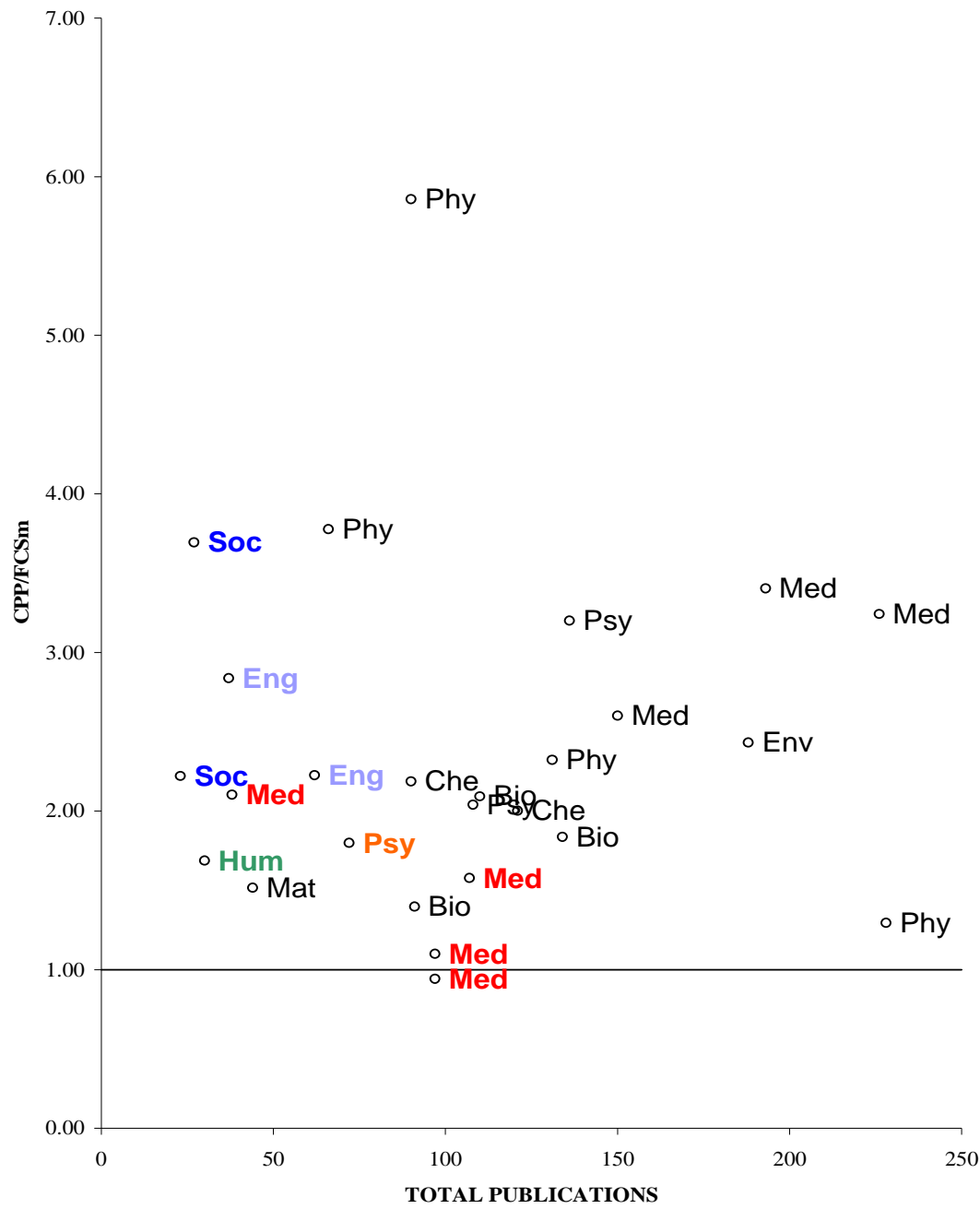
Examples of Hirsch-index values



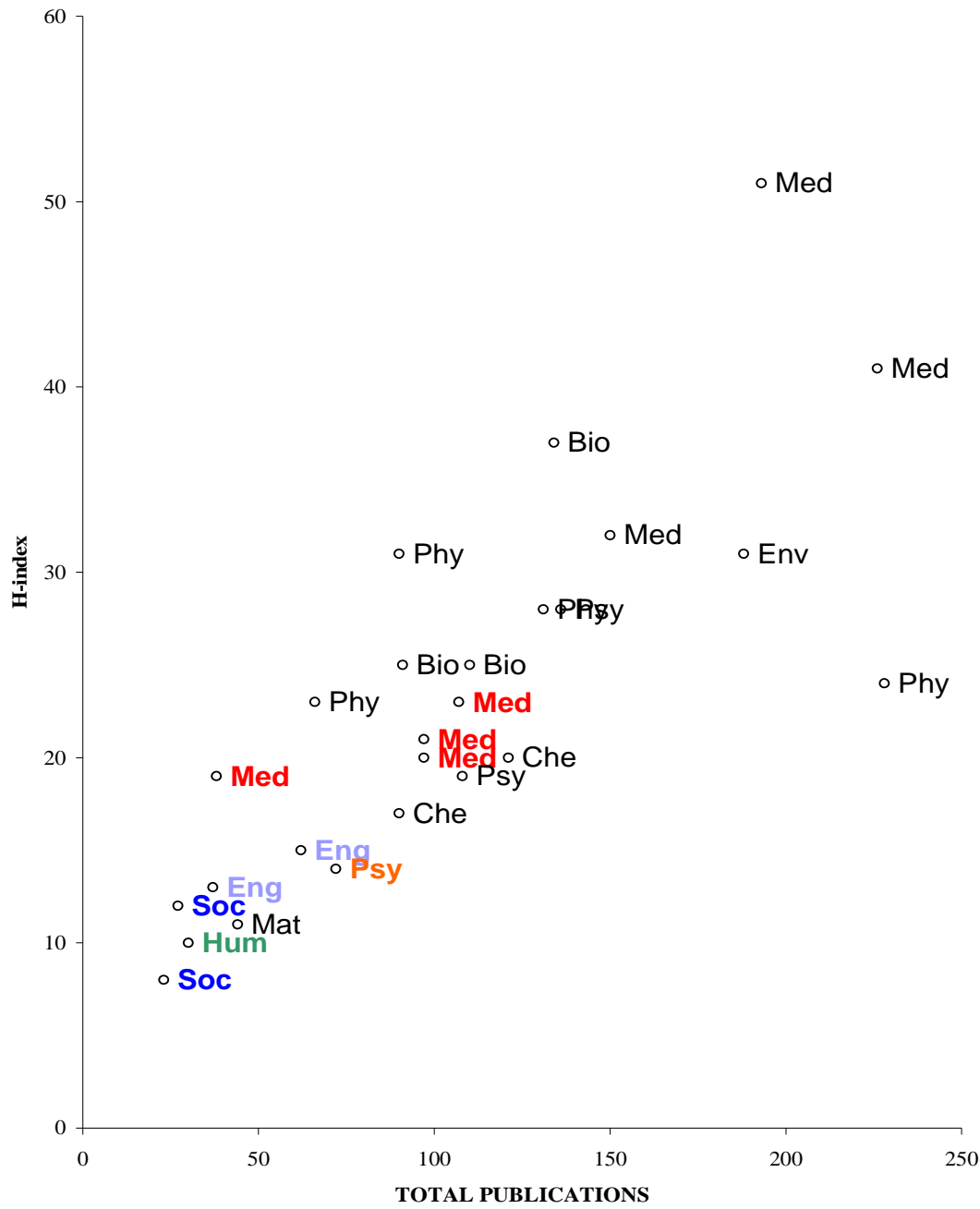
- Environmental biologist, output of 188 papers, cited 4,788 times in the period 80-04.
- Hirsch-index value of **31**



- Clinical psychologist, output of 72 papers, cited 760 times in the period 80-04.
- Hirsch-index value of **14**



- Actual versus field normalized impact (CPP/FCSm) displayed against the output.
- Large output can be combined with a relatively low impact



- H-Index displayed against the output.
- Larger output is strongly correlated with a high H-Index value.

Conclusion on the H-Index

- For serious evaluation of scientific performance, the **H-Index** is as indicator not suitable, as the index:
 - Is insensitive to field specific characteristics (e.g., difference in citation cultures between medicine and other disciplines).
 - Does not take into account *age* and *career length* of scientists, nearly by definition does a small oeuvre necessarily lead to a low H-Index value.