Detection of different types of 'talented' researchers in the Life Sciences through bibliometric indicators: methodological outline

Rodrigo Costas and Ed Noyons

CWTS Working Paper Series

Paper number	CWTS-WP-2013-006
Publication date	November 11, 2013
Number of pages	24
Email address corresponding author	noyons@cwts.leidenuniv.nl
Address CWTS	Centre for Science and Technology Studies (CWTS) Leiden University P.O. Box 905 2300 AX Leiden The Netherlands www.cwts.leidenuniv.nl

Universiteit Leiden

CWTS

Centrum voor Wetenschaps- en Technologie Studies

Detection of different types of 'talented' researchers in the Life Sciences through bibliometric indicators: methodological outline[1]

Rodrigo Costas & Ed Noyons

CWTS, Leiden University (the Netherlands)

Abstract

In this report we describe a general bibliometric methodology developed with the aim of mining bibliographic data to detect different types of "talented" researchers. This case study is focused on the Life Sciences at the worldwide level, but with the particular aim of detecting 'Dutch' and 'Belgian' scholars. This methodology involves different steps that are thoroughly described in this report and also some general results are presented. We also present a discussion of the main advantages and limitations of this methodology as well as possible further research developments are set forth.

1. Introduction

Individual researchers are the nuclear agents of the scientific system and this is the reason why is it so important to study their behavior and performance. Research organizations can change or disappear and, although facilitators, they are not the real producer of the new scientific knowledge. These are the scholars that work and do research. Therefore sometimes individuals need to be considered on their own in order to be able to deeply study the production of scientific knowledge. For the previous reason, the individual is also frequently the object of research evaluations and analysis, and in general it is accepted that it is also important to motivate and to encourage individual productivity in order to improve the research system as a whole (Costas, van Leeuwen, & Bordons, 2010). In general, the evaluation of researchers is vital for the improvement of scientific systems, and bibliometric indicators can contribute to inform this process, although this is not completely free of limitations and problems. Most of them will be discussed in this report.

In general, bibliometric indicators must be handled with great care and their limitations should be considered in any study. However, this is even more important at the level of individuals. The use of bibliometric indicators at the individual level has been an important topic in the bibiometric field for many years (e.g. Costas, 2008; Jiménez-Contreras et al, 2011; Larivière, 2012; de Sousa Vieira, 2013) and recently the debate on the use of bibliometric indicators at this level has been revived in the scientometric scientific community (see for example the 'dos and don'ts in individual level bibliometric indicators' recently promoted by Wolfgan Glänzel and Paul Wouters).

One of the main arguments to take into account in this debate about the use of bibliometrics at the individual level is that science is a multidimensional activity and it should not be evaluated through just one indicator[2]. For this

[1] This study was funded by the Crucell Vaccine Institute, Janssen Center of Excellence for Immunoprophylaxis, Johnson & Johnson.

reason, in our research we propose a combination of several indicators and we support the recommendation of using them with expert judgments (van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003).

At the level of individuals, it is also important to pay attention to the underlying data, since small losses of information may have large influence on the final results, particularly when comparing individuals on the basis of raw indicators. Problems related with the lack of normalization of data in the databases (mistakes in the references, duplications, homonyms among the authors, etc.) can cause problems and need to be understood properly to consider bibliometrics at the level of individuals (Angelo, 2011; Costas & Bordons, 2005). In addition to the previous, probably one of the most important important problems in the application of bibliometric indicators to the analysis of individuals is the low reliability of statistics at this level. This basically means that even with completely accurate counting of publications and citations, small differences in the values are not necessarily meaningful due to the fact of the 'noisy' meaning of citations with small numbers. For this reason, we have a problem of 'uncertainty' in the meaning of the indicators at the individual level. Thus, relying on raw values to compare individuals is an important limitation of bibliometric indicators (but not only, e.g. altmetric indicators or even peer review scores would suffer from the same problem) when applied to the evaluation of individual scholars.

In spite of the above mentioned concerns, individual-level bibliometrics can significantly help to understand better the scientific landscape, and therefore its development is important. In the development and application of bibliometric indicators for the analysis of individual scholars, there are two main analytical perspectives. On the one hand, a more *descriptive* perspective devoted to the detection and analysis of the main individual actors, as well as the aspects that characterize their scientific performance (e.g. 'who are the scholars active in a scientific field or discipline?', 'where do they come from?', 'how are they collaborating?', 'how does their production evolve over time?', etc.); and on the other hand there is a more *evaluative* perspective with more assessment purposes ('who are the most productive scholars in a field?', 'who is publishing in the most impact journals?', etc.) as well as practical ones (e.g., 'who could be a suitable candidate for my organization?'). Very frequently these two perspectives (descriptive and evaluative) interact, and sometimes the descriptive and evaluative perspective go hand in hand. This is the case of this report, as it combines in the descriptive approach (who are the scholars active in the Life Sciences worldwide and how may come from the Netherlands or Belgium?) with a more evaluative one (who and how many are the researchers with a given bibliometric performance level?). With this report we seek to contribute to the debate about the use of bibliometric indicators at the individual by providing a first broad development that can help to detect active researchers in a given scientific domain and also to characterize their different types of bibliometric performance from a more evaluative point of view, considering the most important limitations in the use of bibliometric indicators at this level.

The rest of the report is organized as follows. Chapter 2 presents the main objectives of the report, chapter 3 describes the methodology developed in detail, chapter 4 presents the main and general results of the analysis of the data and information obtained through the methodological development. Finally, chapter 5 discusses the main advantages and limitations of this methodology also sketching some future possible improvements for this methodology.

[2] For example, the h-index is one of the most popular 'single indicators' for measuring individuals but is very inappropriate from different perspectives. It has been proved to be an inconsisten indicator (cf. Waltman & van Eck, 2012) and also to be on the detriment of "selective scholars" (Costas & Bordons, 2007) and younger researchers (van Leeuwen, 2008).

2. Objectives

The main objective of this report is to describe a new bibliometric methodology that has been developed at CWTS in order to detect different types of performances (or "talented") researchers (from a bibliometric point of view) in the field of Life Sciences, with a 'Dutch/Belgian' origin or strong linkages with an affiliation in these countries during their scientific careers.

In order to achieve this main objective, three concrete sub-objectives are targeted:

1) To develop a methodology in order to mine bibliographic data from the Web of Science that allows detecting active scholars in the Life Sciences.
2) To detect as many as possible Dutch/Belgian scholars, either by origin or by showing a strong linkage with any research institutions in the Netherlands or Belgium.
3) To bibliometrically model different types of scientific performance, based on the combination of different bibliometric indicators and the position of the different scholars in the international scientific landscape, thus being possible to create typologies of researchers with different types of bibliometric performance (or 'talents').

It is important to highlight that he proposed methodolgy in this study is probably one the first ones in which a huge sample of Life Science[3] (LS) researchers is considered for analysis and that bibliometric indicators are considered in such a broad scale.

[3] As it will be explained later on in the methodology, this is just a definition of the Life Sciences but not an absolute definition of the field.

3. Methodology

In this section we present the main data sources, conceptual assumptions, indicators and methodological steps we considered in the development of this methodology. The methodology is rather detailed to clearly describe the different steps that have been followed but also to demonstrate the complexity and sophistication of the methodology. In some steps, given the technical complexity of the explanation, only general lines are presented, and future publications on the topic will provide more detailed explanations about them.

3.1. Data sources and periods of time

The database used for the development of this methodology is the in-house CWTS version of the Web of Science database (not including the Conference Proceedings Index). The analysis covers all publications and authors from 1980 to 2011. Citation indicators are calculated over the period 1980 to 2012, hence including one extra year for citations.

3.2. Conceptual assumptions

In the methodological approach we adopt the following starting points elaborated into conceptual assumptions:

- We consider only Web of Science publications. This means that other scientific outputs are not considered in the detection and analysis of the different scholars. No external sources have been considered and therefore all analysis must be assumed as be limited by the data source employed. However, in the fields of Life Sciences, we consider the WoS to sufficiently cover the important scientific output of the most important scholars active in the field.
- Limitations in the data. Bibliographic meta-data has been used in order to describe the different scholars. In all cases, this has been done algorithmically and possible mistakes from the database (e.g. wrong e-mail inclusion, wrong linkages of authors to addresses, incomplete fields, typos, lack of information, etc.) need to be taken into account and some degree of inaccuracy/incompleteness in the results must be contemplated when analyzing the results. In any case, we assume that these omissions are relatively minor (particularly in the broad framework considered in this study) and therefore should not undermine our analysis and outcomes.

3.3. Methodological steps

In this section we describe the main methodological steps implemented to render the main results and to meet the proposed objectives:

Step 1. Field delineation of the Life Sciences (LS core) in the Web of Science

The first step consists on the delineation of the "Life Sciences" within the entire Web of Science. In order to perform this step, we used a recently developed classification of publications classification (Waltman & Eck, 2013) (http://www.ludowaltman.nl/classification_system/).

This classification is paper-based. This means that every publication is classified in only single field at each level. In this classification scheme we discern three levels: the top level of 21 main fields (macro-classification), the intermediate level of 784 fields (meso-classification) and the bottom level of around 22,000 topics (micro-classification). The classification used for this study covers the period 1993 to 2011. In this study the intermediate level of the classification (i.e. the meso-classification of 784 fields) is used (hereafter referred to as the 'meso-fields'). This meso-classification is composed by 784 meso-fields and their main outline in terms of main disciplines

is presented in Figure 1. The map shows the citing relations among the 784 meso-fields. The map uses two dimensions to represent their relations. The closer two meso-fields are, the stronger their citation relation. As only the distance among the meso-fields matters, we can rotate or mirror the map as much as we like, as long as the distances remain the same. A clustering of meso-fields is added to enhance the interpretation of the map. The clustered meso-fields represent major disciplines in the entire science landscape.

Figure 1. Map of all 784 meso-fields classification (period 1993-2011)

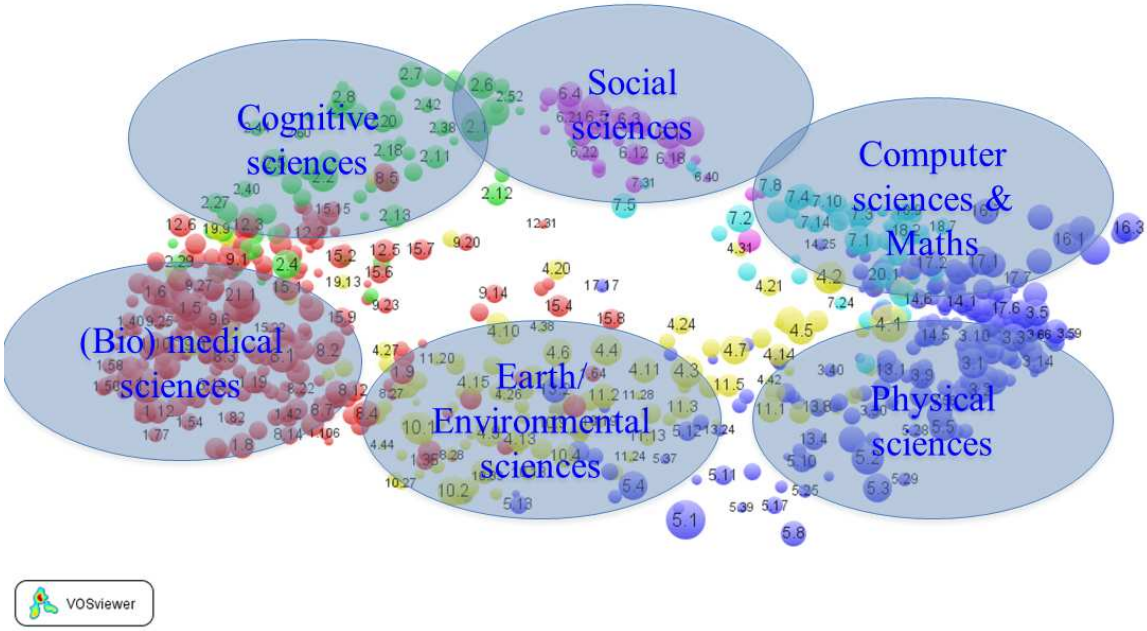
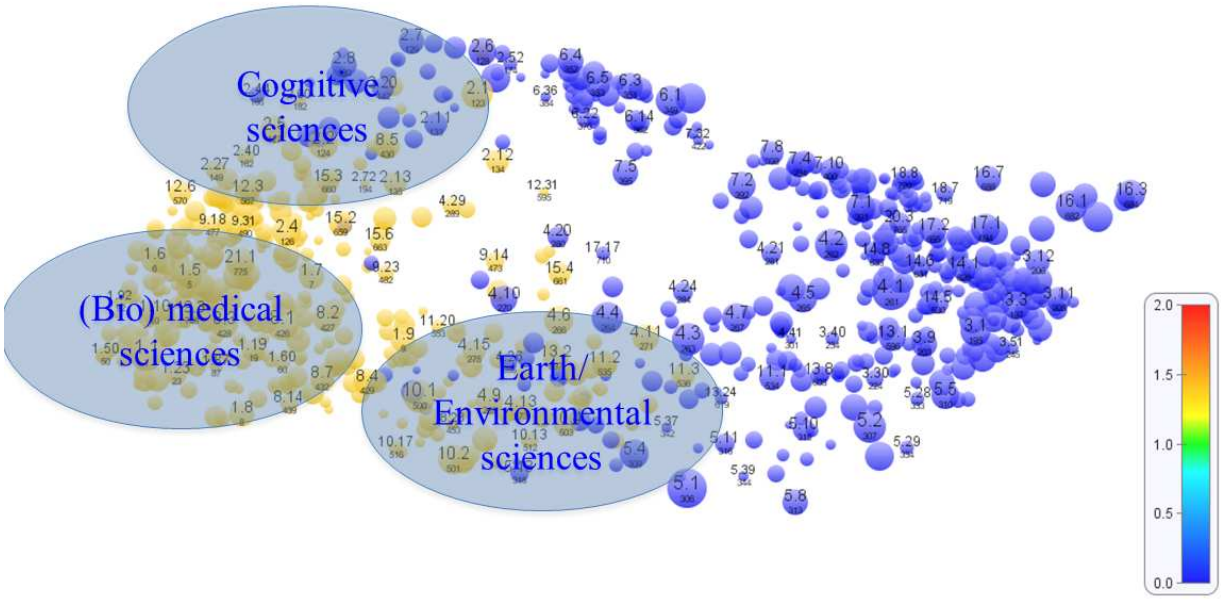


Figure 1 shows that the meso-fields associated with the Life Sciences are on the left-hand side of the figure, dealing mainly with (Bio) medical sciences, the cognitive sciences and to some degree also with the Earth and Environmental Sciences.

Two experts from the Crucell Vacinne Insitute went through the whole meso-classification (i.e. the 784 meso-fields) selecting those that they considered to have relevance for the delineation of the Life Sciences domain. A total of 373 meso-fields were finally selected. Their publications were considered to represent the Life Sciences core ('LS core'). Figure 2 presents the results of that selection. The yellow meso-fields are the ones selected by the two experts to represent the LS core. As it can be seen, the fields that conform the LS core belong to the three main fields previously mentioned, having the main concentration in the (bio) medical sciences. The blue meso-fields positioned among yellow and the yellow ones positioned among blue ones were the discussed in more detail before taking a final definition of the LS core.

Figure 2. Meso fields selected for the creation of the LS core



As a result fo the selection of the meso-fields, more than 8 million publications have been considered to represent the LS core, thus forming a solid base for the detection of all the scholars active in the Life Sciences.

Step 2. Identification of active researchers in the LS core and their full oeuvres

This step consists basically of detecting all individual scientists who have been at some point active in the LS core set of publications. This is one of the most difficult tasks of the methodology as it has to deal with problems such as the homonymy and synonymy of author names (Angelo, 2011; Costas & Bordons, 2007; Wang et al., 2011). In order to minimize this problem, we have profited from an algorithm for author-name disambiguation developed and carried out at CWTS (Caron & van Eck, 2013). This algorithm, which has proven to be very accurate (however not perfect) in the detection of the oeuvres of the authors, has been applied to all the different author name included in the Web of Science data base from the period 1980 to 2012. Documentation on this new algorithm will be published soon (Caron & van Eck, 2013) although preliminary test have shown values of 95% precision and 90% recall.

Step 3. Filtering of scholars active in the LS core

Based on step 2, over 10 million individuals active in the LS-core were detected. We reduced[4] this huge amount of scholars to be included in the analysis by applying several filters.

[4] More than 60% of all the scholars identified have only 1 publication in the LS core.

Step 3.1. Selection of those with ≥5 publications in the LS core

In the first step, we selected only those scholars that have at least 5 publications in the LS core. This filter ensures that we are working only with scholars that have a minimum meaningful scientific production activity in the Life Sciences. By this selection we focus on the 14% of scholars with the highest activity in the LS core (see more details in table 1).

Step 3.2 Selection of those with ≥ 50% of their output in the LS core

This second filter regards the selection of those researchers having at least 50% of their production in the LS core. This filter selects only those researchers having a substantial part (i.e. the majority) of their scientific production in the Life Sciences, hence excluding scholars with only occasional or accidental papers in LS.

For this step, we collected the full oeuvres of the scholars selected in step 3.1. Thus it was possible to calculate the output share of a scholar in the LS core. Hence, we focus on the scholars whose main focus is on the LS, although their full oeuvres will be considered for the subsequent analysis.

Step 4. 'Dutchness/Belgiumness' identification

One of the objectives of the project is also to be able to detect scholars that are considered to have a Dutch/Belgian linkage. This is a very difficult step as here we try to detect scholars not only that have worked in the Netherlands or Belgium but also those that may have a Dutch/Belgian origin. In order to solve this problem we used the full Web of Science database as a demographic database in order to algorithmically detect the potential origin of all the surnames of the database. The execution of this step involves several sub-steps that were applied to all Web of Science data from 1980 onwards.

Step 4.1. Trusted author-country linkages

For all the surnames in all the publications covered in the Web of Science we detected all the *trusted linkages* with countries. By a *trusted linkage* we mean a surname-country relationship that is unambiguously registered in a paper[5]. This step aims at creating all possible and proper linkages between authors and countries according to bibliographic data in the Web of Science. For this purpose we focus only on trusted linkages between authors and countries. This means that only in those cases that there is strong evidence that an author is linked to a country, the link is created and the combination is taken into consideration for the statistical analysis. Steps 4.2 to 4.5 describe these trusted 'author-country' linkages.

Step 4.2. Author-country linkage through the Reprint Author provided in the Web of Science.

The Web of Science contains a bibliographic field where the reprint (or corresponding) author of a publication is directly linked to his/her affiliation (Costas & Iribarren-Maestro, 2007). Using this information it is possible to get another set of *trusted linkages* between authors and affiliations (and countries).

Step 4.3. Registered combinations of author and affiliations contained in the database

[5] It is important to keep in mind that for most publications in the Web of Science, not all the authors are directly linked to their affiliations in the paper, therefore it is very difficult to establish to which affiliation (and country) belongs every author.

From 2008 onwards Thomson has started to record the linkage between authors and countries as they appear in the WoS publications (some previous publications also show this linkage, but these data are more reliable from 2008 onwards). Taking advantage of such a linkage it is possible to create combinations of authors, countries and publications.

Step 4.4. First author - First country combinations.

In a previous study (Calero, Buter, Cabello Valdés, & Noyons, 2006) it was shown that the linkage of the first author with the first address of the publication is quite reliable. Therefore, we considered it also a *trusted linkage* to attribute the first author of the publcations to the first affilation.

Step 4.5. All authors – only one country combinations.

In those publications where there is only one country (although maybe more than one affiliation) all the authors of the publication can be linked to this country. This type of linkage is somehow different from the others, because in this case we do not link an author to a concrete address but only to a country. In other words, we know the country of the author but we do not always know which one of the affiliations is his/hers. This is the reason why in some occasions we know that a scholar had a linkage with the Netherlands or Belgium but we don't know exactly in which organization.

Step 4.6. Calculation of the Kullback-Leibler divergence measure and a normalized Gini Index for all the surnames in the Web of Science database[6]*.*

This step focuses basically in trying to determine the country of origin of a given surname present in the WoS database. The Kullback-Leibler (KL) divergence was applied in bibliometric research for other purposes (Torres-Salinas, Rodríguez-Sánchez, Robinson-García, Fdez-Valdivia, & García, 2013). Particularly in this paper we use it in a similar way as by Waltman & van Eck, who used this indicator to determine the international/national orientation of scientific journals based on the distribution of countries of the authors of the publications. In this case, we use exactly the same approach, but instead of scientific journals we consider surnames.

Besides determining the international orientation/localness of the surnames we also need to estimate the country to which a surname is most probably linked. For this we took a very straightforward approach: the country with the highest number of publications for a surname is considered to be the most 'natural' of that surname.

A problem that we may face with this previous KL approach is that if we take a surname that appears 10 times in country A and 9 times in country B the KL can still be very high, but in fact the possibilities of the surname to be from the second country are also very high. To estimate better this point, we calculated a normalized Gini index (Carpenter, 1979). This index gives an indication on the concentration of a surname over countries. Thus a surname with a strong concentration in one country would have a high Gini index, while a country that is very evenly distributed over countries would have a lower Gini index. Thus, although we can find surnames that belong to only a few countries (and thus having a high KL divergence), we can measure as to how concentrated they are in only one country (thus discarding the possibility of this being natural of another country).

Finally, in order to determine if a surname is 'Dutch' or 'Belgian' we calculated the percentiles of all the surnames in the database by the KL and Gini index. In those cases that the surname has the Netherlands or Belgium as the most important/probable country and the surname is within the Percentile 80 of the distribution of both measures

[6] We expect to publish this methodology in more detail in the future.

internationally. In those cases, authors with a surname falling within this threshold are considered to have a 'Dutch' or 'Belgian' origin although their affiliation may be outside the Netherlands or Belgium.

We realize that this methodology is quite exploratory and novel. Therefore, more research, tests and potential refinements will be developed in the future. Still, all the manual checks of surnames of our algorithms (e.g. Wikipedia Dutch and Belgian surnames, other surnames known by the authors, etc.) proved that the approach yields reliable results so far.

Step 4.7. Final identification of 'Dutch'/'Belgian' scholars

Based on all the previous sub-steps it is possible to identify those scholars that have a Dutch or Belgian linkage. This identification is based on two main criteria:

1. The scholar has a 'Dutch/Belgian surname' (based on the KL – Gini index algorithms), or;

2. The scholar has at least 10% of his/her publications with a 'trusted linkage' with a Dutch/Belgian affiliation (based on all the author-country 'trusted linkages' created in previous steps).

Step 5. Identification of the meta-data for all the individuals finally selected

Apart from detecting individual scholars active in the LS core, it was also important to find with some degree of certainty, other personal information elements that can help to better identify who these scholars are. For this identification we used all the meta-data available through the Web of Science database. The elements considered for the identification of the scholars are the following:

- Author name: among all the author names (including surname and initials, e.g. 'Goudsmit, J') that were attributed to the scholar we chose the one that appears in most of the publications (e.g. if an author appears 20 times with 'de Jong, JB' and 2 times as 'de Jong, B', we will chose the first one. In case of ties, we chose the one with most characters).
- First author name[7]: as for the author name, we selected the most frequent first name of the author, and in case of a tie we chose the longest one in terms of number of characters.
- E-mail[8]: whenever available, we detected the most common e-mail address of the scholar. If different e-mails were found across years, we selected the most recent one. In case of ties, again the variant with more characters is selected.
- Most Common Address (MCAD): based on all the previously explained 'trusted linkages' between an author and the affiliations in the publication, we detected the most common address (in terms of number of publications) of every author. If there are no *trusted linkages* between a scholar and any affiliation, this field is missing.
- Most Probable Recent Address (MPRAD): this is basically the MCAD of every scholar in the last year of publication. Thus we were able to estimate where the scholar has been working in the most recent period.

[7] Web of Science occasionally includes the full first name of the author. This is not available for all the authors but we have used it whenever it was available.

[8] E-mail data is also occasionally available in the Web of Science, particularly for reprint authors. This means that not in all cases we have been able to detect an e-mail for every scholar identified.

Step 6. Calculation of standard bibliometric indicators and age-related indicators

For all the relevant scholars active in the Life Sciences (i.e. the 1,309,458 set of scholars finally identified) we calculated standard bibliometric indicators based on their full oeuvres (i.e., their LS-core publications + publications outside the core). The period of time for publication spans from 1980 to 2011; citations up to 2012. Only article, reviews and letters[9] were considered. For a description of the CWTS standard bibliometric indicators, see Waltman, Van Eck, Van Leeuwen, Visser, & Van Raan (2010); Waltman, van Eck, van Leeuwen, Visser, & van Raan (2011).

For every scholar identified we calculated the following indicators:

- *P*: total number of publications
- *TCS*: total number of citations (excluding self-citations)
- *MCS*: mean citation score.
- *MNCS*: mean field[10] normalized citation score
- *TNCS*: total normalized citation score (P*MNCS)
- *P_uncited* and *PP_uncited*: number of publications not cited and its proportion (pp)
- *Ptop10* and *PPtop10*: number of publication within the top 10% of the publications in the same field(s) and their proportion.
- *MNJS*: mean normalized journal score
- *h*-index: following the methodology by Hirsch (2005). For this indicator we have considered all outputs (not only articles, reviews and letters).

In addition to the previous indicators and based on the 'full oeuvres' of the scholars detected we were also able to determine 'age' related indicators. They involve the following:

- First year of publication: first year when the scholar published his/her first publication. This is useful to create cohorts based on scientific age.
- Last year of publication: last year when the scholar published his/her last publication.
- Scientific life: difference between the last and the first year of publication. It is an estimation of the period of scientific life of the scholar (the scientific age).

Step 7. Modeling 'scientific performance'

In this step 'scientific performance' in terms of bibliometric indicators is modeled. This modeling is inspired by a similar approach previously developed at CWTS (Costas et al., 2010). In this approach, bibliometric indicators are related to three different dimensions of performance that can be captured through bibliometric indicators:

- Total performance of total production dimension. Here we have the size dependent indicators both of production and impact (e.g. *P*, *TCS*, or the *h-index*). We selected the indicator P as being the most descriptive of the overall production of the scholars. This dimension measures the production capacity of a scholar.
- Average impact dimension. These are the indicators that measure the average impact of the publications of the scholars, both normalized or non-normalized (e.g. *MCS*, *MNCS*, or *PPtop10*). In this case we selected the *PPtop10* indicator as the most representative of this dimension and also the less sensitive to outlier

[9] Letters are weighted 0.25, where normal articles and reviews are weighted 1
[10] Here 'fields' are the Web of Science Journal Subject Categories.

publications (e.g., isolated publications with a very disproportionate number of citation could influence indicators such as *MCS* or *MNCS*, but less the *PPtop10* indicator). This dimension describes the ability of the scholars to keep a sustainable high level of impact in most of his/her publications and not only on a few of them.

- Journal quality impact. This dimension refers to the impact of the journals in which scholars publish (e.g., the journal impact factor-like indicators or our *MNJS*). In a way, this dimension refers to the ability of a scholar to place his/her publications in high visible journals. We considered the *MNJS* indicator as the best representative of this dimension.

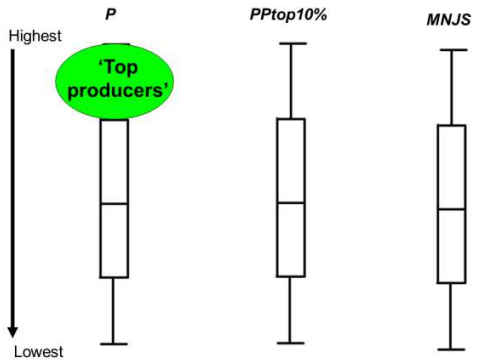
Taking into account these three indicators per scholar and considering all the scholars in the LS worldwide (not only the Dutch/Belgian scholars), we calculated the top percentiles 25 and 50 for distribution of the 3 indicators across all the scholars worldwide. Thus, based on the presence of the individuals across the different percentiles it is possible to classify individual scholars depending on their performance in these three dimensions, contextualized by the all other scholars worldwide (cf. Costas et al, 2011).

Based on this approach, we identified three types of bibliometric performers:

1) Top producers

These are all the scholars that are within the top 25 percentile of the distribution of researchers by the indicator P. In other words, they are among the 25% most productive LS researchers worldwide (see figure 3)

Figure 3. Definition of 'Top producers'

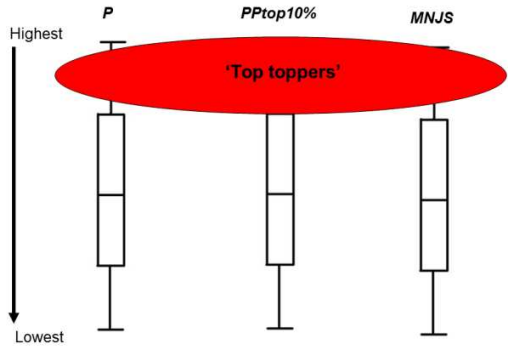


The advantage of this type of performers is that it is very simple to explain and to understand. On top of that it is in line with the general perception of performance (e.g. it is strongly related to the same group that we would detect by ranking the scholars by the h-index). The main limitation is also that it is too simplistic. It is primarily one dimensional and easy to manipulate. The number of publication says nothing about the quality of the papers. The ability of the scholars to produce high impact publication and in high impact journals is not recognized. For this reason, we have the second type.

2) Top toppers

This is a more competitive type of performance. Basically to be a top topper, the scholar must be within the percentile 25 of production (i.e., he/she must be a top producer) but also within the percentile 25 of the dimension of average impact (i.e. the PPtop10%) and among the percentile 25 of the journal impact quality (i.e. the MNJS indicator). In other words, they must be among the 25% most productive individuals, among the 25% of the best producers of highly cited publications and among the 25% of the best publishers in higher impact scientific journals (see Figure 4).

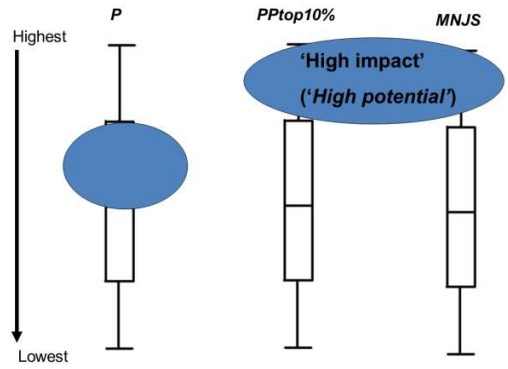
Figure 4. Definition of 'Top toppers'



3) High impact / high potential

Clearly, top toppers represent a very competitive type of researchers. However, we could still foresee a type of scholars who, without having published the same amount of publications as the top toppers (or the top producers), still would represent an interesting type of performers. These are the scholars that have a more 'selective' publication strategy (cf. Costas & Bordons, 2005; Rodrigo Costas et al., 2010) or that are more 'perfectionist' in their work (publishing slightly less publication than other colleagues, but still of high impact and in very good journals). These performers are defined as presented in Figure 5. In essence, in our modeling they must have a level of production between the percentile 75 and 50 of the most productive of the world. In order words, they still need to have some level of production, although not that high as to be a top producer. Simultaneously they also must have a high share of highly cited publications (i.e. they must be among the 25% of the world in the PPtop10% indicator) and publish in high impact journals (i.e. among the 25% of the world in terms of the MNJS indicator).

Figure 5. Definition of 'High impact' or 'High potential'



In this group we can also distinguish 2 types of performers, depending on the time when they have been active. In this sense, if they have been active before the year 2000, they are considered as being 'high impact', while if they belong to the most recent cohort (i.e. they started to publish between 2000 and 2011) then they are considered as 'high potential', due to their relatively recent incorporation to the scientific publishing community.

It is important to mention that these three typologies are not the only ones that can be extracted from our methodology. Other typologies could be also possible to define (e.g. in Costas & Bordons (2008) other typologies as 'big producers' or 'low producers" were also envisioned) but for the purpose of this report these have been considered the most relevant typologies of bibliometric performance. In any case, clearly the methodology opens the door to the exploration of other types of typologies.

Step 6. Final selection (250 researchers and validation list)

Based on the previous modeling we selected the final list of candidates based on the following criteria. This is in order to present a more detailed analysis of a selection of scholars. For this selection we have applied the following filters:[11]

1. Only researchers within the P25 of P (i.e. scholars among the 25% most productive LS researchers in the world), within the P25 of MNJS (i.e. scholars among the 25% scholars publishing in the best journals) and within the P50 of PPtop10% (i.e. scholars among the 50% of scholars that have the highest shares of Ptop10% publications).
    a. Top 150 researchers sorted by Ptop10% (i.e. the total number of highly cited publications of the scholars) from the full period with a Dutch/Belgian origin/linkage.
    b. Top 100 researchers sorted by Ptop10% from the most recent cohort with a Dutch/Belgian origin/linkage.

[11] It is important to note that we do not assume any normative qualitative value in this selection. This selection was just in order to get a better and deeper view of the finally identified scholars, but we do not assume that this is the "best" selection that could have been done.

4. Main results

In this section we outline the main results regarding the data collection and identification of scholars, as well as the main results regarding the different types of performance previously discussed. The results only involve numbers. In this report we will not discuss individuals as such and our main objective is more to exemplify other new types of possible results that this type of development would allow in bibliometric studies at the individual level.

Main figures in the selection of scholars

Table 1 presents the summary of the main data processed during the development of the project.

Table 1. Main figures regarding the identification of scholars active in the Life Sciences

Data	#
Scholars in LS worldwide	10,008,311
Scholars in LS ≥ 5 pubs	1,388,080
Scholars ≥ 50% in LS *	1,309,458
Dutch/Belgian scholars in final set	58,281

* *All percentile calculations are based on this set!*

We estimate that more than 10 million scholars have published at least one publication in the Life Sciences in the period 1980-2011. However, only 1,388,080 (14%) have 5 or more publications in the field. After collecting their full oeuvres (i.e. the outputs for every scholar within and outside the LS-core) we discarded 78,622 cases of scholars that didn't meet the requirement of having at least 50% of their output in the LS-core. As a result a total of 1,309,458 individual scholars were identified and this is the list of scholars that was used for the calculation of percentiles and the modeling of the different types of bibliometric performance.

Main values regarding Dutch / Belgian scholars

In this section we focus on those researchers with a Dutch or Belgian origin or a strong link with any of the two countries. In total we found a total of 58,281 cases representing around 5% of all the scholars in the final population of LS active scientists worldwide. In the following tables (tables 2 and 3) we present the main results regarding the affiliation linkage of these scholars considering their Most Common Address (MCAD) and their Most Probable Recent Address (MPRAD). It is important to know that through Web of Science data it is not always possible to attribute the exact affiliation to an author. This is the reason why there is a number of cases for which we cannot attribute a proper country to the authors (i.e. row "'Dutch/Belgian' scholars with a NL/BE linkage, but not certainly defined").

Table 2. 'Dutch' or 'Belgian' scholars active in the Life Sciences and affiliation based on the MCAD

Data	#	%
Scholars with their MCAD in NL	26,083	45
Scholars with their MCAD in BE	12,008	21
'Dutch/Belgian' scholars outside NL or BE[12]	11,022	19
'Dutch/Belgian' scholars with a NL/BE linkage, but not certainly defined[13]	9,168	16
Total	58,281	

In Table 2 we can see how in total 26,083 identified individuals (45%) have the Netherlands as their most common address, 12,008 (21%) have an affiliation in Belgium. For 11,002 individuals we found that they have a Dutch/Belgian surname and/or were inked at least in 10% of their papers to the Netherlands or Belgium, but their most common affiliation is outside the Netherlands or Belgium. In other words, either they have a Dutch/Belgian surname or they lived/worked at some point in any of these two countries but their current affiliation is abroad. Finally 9,168 scholars have a Dutch/Belgian surname (or were linked to the country in at least 10% of their publications) but we were not able to attribute a definite affiliation (nor in NL/BE or abroad). Therefore, they don't have a proper address. This is true both for the MCAD and the MPRAD.

Table 3. 'Dutch' or 'Belgian' scholars active in the Life Sciences and affiliation based on the MPRAD

Data	#	%
Scholars with their MPRAD in NL	25,728	44
Scholars with their MPRAD in BE	11,918	20
'Dutch/Belgian' scholars outside NL or BE[14]	11,467	20
'Dutch/Belgian' scholars with a NL/BE linkage, but not certainly defined[15]	9,168	16
Total	58,281	

Regarding the Most Probable Recent Address of the scholars, the main results are presented in Table 3. The values in the two tables (2 and 3) show relatively small differences. Roughly the groups have the same size in MCAD and MPRAD. We calculated that the number of researchers abroad (i.e. the number of "'Dutch/Belgian' scholars outside NL or BE") has slightly increased by 400 cases. For now, it is early to conclude that this indicates a decreasing mobility of researchers in the Netherlands or Belgium. It needs to be investigated in more detail first.

[12] Researchers with a Dutch or Belgian surname but their MCAD outside the Netherlands or Belgium.
[13] Researchers with a Dutch or Belgian linkage but for whom is not possible to algorithmically detect their NL/BE MCAD affiliation (e.g. because they do not have enough publications directly linked to a concrete affiliation in these countries, etc.)
[14] Researchers with a Dutch or Belgian surname but their MPRAD outside the Netherlands or Belgium.
[15] Researchers with a Dutch or Belgian linkage but for whom is not possible to algorithmically detect their NL/BE MPRAD affiliation (e.g. because they do not have enough publications directly linked to an affiliation, etc.)

Main values regarding the different typologies of performance – Global analysis

In this section we present the main figures regarding the different performance typologies of the scholars previously identified. Table 4 presents the results for the whole population of worldwide identified scholars in order to present the main figures that later on can work as a reference value.

Table 4. Distribution of top performers for the full list of LS scholars finally identified

	Total scholars	%	Top producers	%	Top toppers	%	High impact	%
Full period	1,309,458	100	327,375	25	61,567	4.7	49,109	3.8
Cohort	614,318	100	153,593	25	25,573	4.2	24,782	4.0

As table 4 shows, not surprisingly we have a 25% of top producers (as they are defined by the percentile 25 of the most productive scholars worldwide) in both periods of time. Focusing on the full period analysis we have that only 4.7% (for the full period) and 4.2% (for the most recent cohort) of all the researchers worldwide qualify as "top toppers". This lower share of researchers as top toppers is not a surprise if we take into account that this is a very 'though' selection of researchers. They have to perform high not only in output, but also in their share of highly cited publications and in the impact of their publishing journals. Finally the "high impact" typology represents around 3.8% of the international scholars and 4% among the most recent cohort. The "High impact" or "High performance" group is an interesting one because is hardly difficult to identify with the most common size-dependent indicators (i.e. h-index, total production, number of highly cited publications, etc.) that normally detect the most productive ones but not those that are moderate in production but whose publications are highly cited.

Main values regarding the different typologies of performance – Dutch and Belgian researchers

In this section we focus on the presence of Dutch and Belgian researchers across the three types of performance previously described. Table 5 presents the results.

Table 5. Distribution of top performers for the Dutch/Belgian LS scholars finally identified in the full period

Group of scholars	Total	%	Top producers	%	Top toppers	%	High impact	%
Total NL/BE identified scholars	58,281	100	15,393	26	2,821	4.8	2,737	4.7
Scholars with their MCAD in NL	26,083	100	7,561	29	1,576	6.0	1,253	4.8
Scholars with their MCAD in BE	12,008	100	3,511	29	486	4.0	410	3.4
Dutch/Belgian' scholars outside NL/BE	11,022	100	2,622	24	606	5.5	601	5.5
Dutch/Belgian' scholars with a NL/BE linkage, but not certainly defined	9,168	100	1,699	19	153	1.7	473	5.2

Table 5 shows the results for the scholars that have a Dutch/Belgian linkage (i.e., 58,281 scholars). Comparing their results to those in table 4, we can see how Dutch/Belgian scholars overall are proportionally more represented

among the Top producers (26% vs. 25%). Also regarding high impact (4.7% vs. 3.8%) they are better represented. The share of top toppers is quite similar as for the whole population (4.8% vs. 4.7%).

If we focus on those with a most common affiliation in the Netherlands (26,083 scholars), we see that they are even better represented not only among the top producers (i.e. 29% vs. 25%), but also among the top toppers (6.0% vs. 4.7%) and among the high impact scholars (4.8% vs. 3.8%). However, if we focus on those scholars with an affiliation in Belgium we can see how in this case they are above the international level in terms of top producers (i.e. 29% vs. 25%) but underrepresented among the top toppers (i.e. 4.0% vs. 4.7%) and among the high impact scholars (i.e. 3.4% vs. 3.8%).

'Dutch/Belgian' scholars with an affiliation outside the Netherlands or Belgium are slightly underrepresented among the top producers (24% vs. 25%) but they are overrepresented among the top toppers (5.5% vs. 4.7%) and even more so among the high impact researchers (5.4% vs. 3.8%).

Finally, those scholars for whom we were not able to attribute a Dutch or Belgian affiliation, are underrepresented among the top producers (19% vs. 25%)[16], and even more among the top toppers (i.e. 1.7% vs. 4.7%). Still, they are overrepresented among the high impact scholars (5.2% vs. 3.8%).

[16] This is actually not a surprise. If we take into account that the linkage of authors-affiliations is based on algorithmic and probabilistic methods, it makes sense that for those who was not possible to find an affiliation their production is relatively lower, being this the reason why it was not possible to find for them a trusted affiliation.

5. Discussion

In this section we discuss the methodology developed, paying special attention on the assets, limitations and future challenges. Furthermore, research lines regarding this methodology are set forth.

5.1. Assets of the methodology developed

The methodology developed regards novel and strong points, particularly compared to developments in the field of bibliometrics so far. To the best of our knowledge there hasn't been similar approaches in analyzing the performance of scholars in the Life Science at this level and particularly not at this scale (worldwide).

In this section we discuss the main assets of the methodology developed.

1. Novel and robust field delineation. For the delineation of this project we have used a novel field classification available at CWTS (Waltman & van Eck, 2013). This novel classification has two strong advantages compared to other approaches:
   - *Paper-based classification*. This means that publications are individually classified in fields, thus avoiding the common problems of classifications based on journals (e.g. Journal Citation Reports classification) that pose the intrinsic problem of considering all publications in a journal to belong to the same field, without this not necessarily being true. Considering that this is a study at the individual level, counting with such classification is a strong advantage because it is also possible to detect authors that are publishing in LS but outside the LS journals.
   - *Very detailed classification*. Another important advantage is that this novel classification is very detailed. It is composed by more than 700 fields, thus making the delineation of the LS core more accurate (as shown in our analysis).
2. Broad scale of the analysis and robustness. This analysis is unique, including elements such as its coverage over time, its scope (all active scholars), number of 'scholars' detected and analyzed (over a million), etc. Although without claiming to be perfect (as it will be explained in the limitations section) we can consider that given the huge amount of data (publications as well as individuals) the results are robust and informative.
3. Novel consideration of a bibliographic database as a demographic databases. In the developed methodology we made an attempt to estimate the country of origin of the surnames of the author in the database. To the best of our knowledge this has been rarely done before. Previous studies (Kissin & Bradley, 2013; Kissin, 2011) focused on other groups (e.g., Israeli researchers), but never on the same scale as in this project, thus proving the possibilities of bibliographic databases also as kind of demographic databases.
4. Construction of indicators based on the mining of bibliographic data to characterize individuals: among these new indicators we have age-related new indicators (e.g. scientific age, number of gap years, etc.), and also the possibility of estimating elements such as the most common affiliation of a scholar (in our terminology MCAD), the most recent affiliation (in our terminology MPRAD), trusted e-mail addresses, etc.
5. Multidimensional approach in the analysis of scientific performance. This is one of the most innovative elements of this report. We considered multiple bibliometric dimensions (particularly size dependent and size independent dimensions) vs. other approaches that are basically size dependent (e.g. h-index). Given this multidimensional approach, we developed a very competitive analysis of the different scholars, in the sense that for researchers to perform high in all dimensions they need to be good not only in terms of numbers of publications, but also regarding their share of highly cited publications and regarding the quality of the journals where they are publishing. This multidimensionality of the approach also suggest the difficulty in its potential manipulability by the individuals.
6. International and contextual performance analysis. Another important innovative element of this methodology is that we benchmark all the scholars internationally and not among small sets of individuals

(e.g. not only the Dutch or Belgian scholars). This is a strong advantage of the methodology because it allows to contextualize the performance of a scholar globally. In essence, it means that if a researcher in a country (e.g. the Netherlands) would have a lower performance compared to his/her compatriots but still high internationally, the methodology would reveal this. Thus, we avoid the narrower perspective that other approaches have focusing on limited sets of researchers (e.g. a single unit, university or country).

7. Percentile approach for the determination of types of performance. This is also an important asset of this methodology. Relying on raw bibliometric scores for the bibliometric analysis of individuals is problematic. In the first place, we have the more technical problem of the data collection and data quality (e.g. it is very easy to omit some publications for a scholar that would represent changes in his raw indicators); secondly the problem of the lower reliability of bibliometric indicators at this level. Therefore raw counts of publications and citations can only be considered proxies of the actual performance values for every author. In order to mitigate this problem of the lack of reliability of bibliometric indicators at this level, the use of percentiles help to better contextualize the scores obtained for the different individual. In essence, if we can expect that the raw values of the indicators of the different scholars can be slightly different (e.g. due to mistakes in the data collection, author name disambiguation, or the more conceptual problem of the meaning of citations) we can fairly expect that a given scholar won't change that much his/her percentile position if we work with relatively broad percentile classes (as done in this methodology). Finally, the use of percentiles also helps to easily discuss the results at the individual level in a lay person language (i.e. the understanding of percentages is easier than sometimes the more complex meaning of individual raw indicators) by being able to make statements such as 'this scholar belongs to the 25% most productive group of LS scholars worldwide as measured by the Web of Science'.

5.2. Limitations of the methodology developed

As discussed in the previous section, the developed methodology has strong advantages making it a unique tool for mining bibliometric data and characterizing the performance of individual scholars from a bibliometric perspective. However, this methodology is not completely free of problems and still poses some limitations that need to be taken into account when considering the results.

1. Only bibliometric performance as covered in the Web of Science is here presented. This methodology is limited to the Web of Science database which basically is focused on English-language scientific journals. This means that no other outputs (e.g. books, book chapters, articles in local languages, etc.), scientific or not, are considered, and also no other types of impact apart for the citation impact (e.g. social impact, educational impact, health impact, economic impact, etc.) are considered.
2. Only scientific publishing activities are considered. It is important to remark the idea that with this methodology and type of studies we are focusing on only one part of the scientific activities of scholars. This means that other activities such as teaching, training, patenting, funding acquisition, consultancy, media outreach, etc. are not considered at all. Thus it is important to bear in mind that this analysis is only limited to a type of output and activity (i.e. the publication of scientific articles). In addition, we only focus on publications that are covered by the Web of Science, this meaning that other scientific outputs covered in other databases are not considered.
3. Data quality. In this methodology all data and results have been obtained algorithmically (e.g. author identification, author-affiliation/country linkage, characterization of the individuals, etc.). This means that data errors can still happen and that all results and particularly raw values need to be considered with some care and caution. No strong arguments should be made based on raw values of indicators, linkages between authors and addresses, e-mails, etc. In the same line, the algorithms tend to work better when the authors are

productive and do not have a very common name, therefore when dealing with authors with relatively low levels of production or very common names, caution should be higher.

4. Conceptual problems. The methodology here presented also poses several conceptual problems that need to be acknowledged. These problems can be summarized as follows:
   a. *Thresholds*: along the methodology and particularly in the selection of the percentiles there are thresholds that are established (e.g. percentile 25, percentile 50, etc.). These are necessary choices but that are not free of some degree of arbitrariness (e.g. why P25 and not P33?) and still based on them we can make some "unfair" differences (e.g. a scholar just below the percentile 25 of one of the chosen indicator would not qualify as a 'top topper'). Other thresholds are set in some of the steps of the methodology (e.g. 5 publications, 10% of the publications with trusted linkage with the Netherlands or Belgium, etc.), and although they have been conveniently tested (i.e. several tries have been performed with other test, and they normally were considered the best choice), they still could be improved.
   b. *The real age of the scholars is not known*: although we consider the first year of publication of a scholar her scientific 'birth' (i.e. from there onwards we count her 'scientific life'), that does not mean that this is the real age of the researcher and therefore we can consider them only as proxies of the age of the scholars.
   c. *Multidisciplinary scholars*. Another important conceptual problem is that those scholars that have a very multidisciplinary profile (i.e. that they are active in several fields of science apart from the LS-core) could fail in their detection as active scholars in the LS[17].
5. General problems of bibliometrics also apply at the individual level. Finally, it is important to keep in mind that general problems related with bibliometrics also apply at the individual level. Among these problems and particularly important for this project are the following:
   a. *Normalization of indicators*. The normalization of indicators is based on the Journal Citation Reports Subject Categories of the Web of Science. Although convenient (as this brings a common and well known framework for the normalization of the indicators) this normalization is not free of limitations. In the first place, these classifications are based on journals (not on individual articles), secondly the more basic areas of the different disciplines (which normally have higher citation densities) have an advantage compared to clinical research (van Eck et al, 2013).
   b. *Limited value of citation analysis*. Citation analysis as such is not free of limitations, the meaning of citations (e.g. negative citations, perfunctory citations, etc.), the possibilities of manipulation (by scholars or journals) or the determination of the actual contribution of the scholars to the papers (e.g. behavioural problems such as 'Honorary authorship' or 'salami slicing') also (and particularly) apply to the analysis of individual scholars. Therefore general cautions regarding bibliometric analysis also need to be observed when considering the results of bibliometric studies at the individual level.

5.3. Future challenges and research lines

Based on the previous limitations, we can argue that this study opens many opportunities for research on bibliometric analysis at the individual level. Here we summarized some of these research lines, focusing first on that research that is necessary in order to overcome the main limitations of this methodology, and secondly in order to highlight other lines of research that could be developed and/or reinforced with these methodologies.

[17] For example suppose a scholar with 40% of her output in the LS-core, 30% in Chemistry, 20% in Physics and 10% in the Social Sciences, although the scholar has a majority of her output in the LS-core, her production in other fields is higher and therefore it would left out of this study.

Challenges (solutions to the limitations previously presented):

- *Inclusion of other sources (e.g. other databases) and other types of impact (e.g. altmetrics, funding acknowledgments, webometrics, etc.) in the analysis of individual scholars*. This would enrich the analysis and the set of dimensions considered in the analysis of the individuals, thus giving the opportunity to detect other profiles of researchers not detected before, as well as other types of performance (e.g. highly influential scholars in the social media) not studied until now.

- *Data quality*. After the development of the methodology, there are several aspects and steps of the methodology that are subject of continuous improvement. Here we can mention:
    - Improvement of the author disambiguation algorithm.
    - Improvement of author-affiliation linkage algorithm. Clearly, this is something that will improve overtime as databases such as Web of Science or Scopus record better the addresses and the linkage of the authors with their affiliations. Based on the existence of more certain data we could more easily estimate the quality of our algorithm and to improve it accordingly.
    - Improvement of surname origin/linkage detection. An interesting idea would be to study a validated 'golden set' of surnames for which we unambiguously know their origin and test it with our results. Unfortunately, so far, we are not aware of the existence of such validated gold standard.

- *Investigation of the more conceptual problems*. Conceptual problems are the most difficult ones. This is because they are not caused by technical or data-related (only) problems, but by more fundamental issues. Anyway, some research lines targeted to solve these problems are presented below:
    - Improvement of the percentiles approach. An interesting alternative to the percentile approach (which involves the need of choosing some thresholds) could be the introduction of approached more based on 'laws' or general properties of the populations of researchers across disciplines (e.g. the Lotka, Characteristic Scores Scales, etc.). Also, the consideration of bootstrapping techniques in order to better determine the degree with which a scholar belongs to a particular percentile class will be explored in the future.
    - Discovery of other performance dimensions of scholars and their perception. This line consists on the study of other dimensions of performance related to scholars (not possible to capture through bibliometrics) and that can be analyzed through other informetric techniques such as Altmetrics or Webmetrics.
    - Inclusion of other bibliometric elements in the analysis of scholars: collaboration, network analysis and development of new indicators (e.g. fractional countings, etc.) and exploration of other multivariable approaches.

Research lines:

In addition to some of the previous lines of research targeted to improve the methodology and the bibliometric analysis of individuals, we can also point out some lines of research that can be developed or strongly supported by this methodology (and its improvements) and the collection of bibliometric data and indicators at the individual level:

- Investigation of age effects over the productivity of individuals and potential adaptation of indicators/benchmarks accordingly. Also the study of the individual determinants of scientific performance.
- Research on the real possibilities of bibliometrics at the individual level and suggestion of proper uses. This is still an open debate in the scientometric community.
- Gender analysis.
- Mobility studies (including demographics/migrations) and refinement of the algorithm for surname origin.
- Bottom-up and individual bibliometrics evaluations.
    - Are all the scholars or an organization of high level? Outliers? How are the scholars of one country/university compared to another?
    - We can easily analyze the performance of a university based on the publications that carry the address of such university. However, universities 'do not produce research', persons linked to some degree to those universities actually produce the research. Therefore it makes sense that also looking at the human and personal perspective of those units makes sense. How are producing the scholars of a given institution is also of relevance for the whole performance of the university.
    - As shown in this report it is also possible to study the difference of the countries, universities, etc. based on the number of top performers that can be (with some degree of certainty) linked to them. In this sense, an interesting indicator could the share of top toppers of a given country or university, comparing it with the international distribution as a benchmark.

Clearly, the development of individual level bibliometric studies has a strong potential and future not only for research evaluations but also for a better understanding of the scientific process and the generation of new knowledge. However, we don't want to finish this report without recommending once more the need of paying special care and caution when working with bibliometrics at the individual level. In this case, the reading of publications such as Bornmann & Marx, (n.d.); Costas & Bordons (2005); Costas et al. (2010) and particularly the new proposal of *'Do's and don'ts in individual level bibliometrics'* (http://www.slideshare.net/paulwouters1/issi2013-wg-pw) recently promoted by the Scientometric community, are strongly recommended. In all cases, common sense are expected from the users of bibliometric results at the individual level.

Acknowledgments

This work has been funded by the Crucell Vaccine Institue. The authors of this report are also very grateful to Jessica Meijer, Dick den Os, Jaco Klap and Jaap Goudsmit from Crucell Vaccine Institute, Center of Excellence for Immunoprophylaxis, Johnson & Johnson for all their feedback, comments, support and suggestions during the whole development of this research.

References

Angelo, C. A. D. (2011). A Heuristic Approach to Author Name Disambiguation in Bibliometrics Databases for Large-Scale Research Assessments. *Journal of the American Society for Information Science and Technology*, 62(2), 257–269. doi:10.1002/asi

Bornmann, L., & Marx, W. (n.d.). Standards for the application of bibliometrics in the evaluation of individual researchers working in the natural sciences, (0), 1–34.

Calero, C., Buter, R., Cabello Valdés, C., & Noyons, E. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics*, 66(2), 365–376.

Caron, EAM & van Eck, NJP (2013). Large scale author name disambiguation using rule-based scoring and clustering. CWTS working papers

Carpenter, M. P. (1979). Similarity of Pratt's Measure of Class Concentration. *Journal of the American Society for Information Science*, 30(2), 108–110.

Costas, R. (2008). *Análisis bibliométrico de la actividad científica de los investigadores del CSIC en tres áreas: Biología y Biomedicina, Ciencia de Materiales y Recursos Naturales. Una aproximación metodológica a nivel micro (Web of Science, 1994-2004)*. [Doctoral Thesis]. Getafe: Univesrity Carlos III of Madrid. <http://orff.uc3m.es/bitstream/handle/10016/4947/Rodrigo%20Costas%20Tesis.pdf;jsessionid=B5BA7D967A9B743C327DE9A5FFA66D83?sequence=1>

Costas, R. & Bordons, M. (2007). Algoritmos para solventar la falta de normalización de nombres de autor en los estudios bibliométricos. *Investigación Bibliotecológica*, 21(42), 13–32. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_serial&pid=0187-358X&lng=en&nrm=iso

Costas, R., & Bordons, M. (2005). Bibliometric indicators at the micro-level: some results in the area of natural resources at the Spanish CSIC. *Research Evaluation*, 14(2), 110–120. doi:10.1021/ie034164a

Costas, R., & Iribarren-Maestro, I. (2007). Variations in content and format of ISI databases in their different versions: The case of the Science Citation Index in CD-ROM and the Web of Science. *Scientometrics*, 72(2), 167–183. doi:10.1007/s11192-007-1589-z

Costas, R., Leeuwen, T. N. Van, & Bordons, M. (2010). A Bibliometric Classificatory Approach for the Study and Assessment of Research Performance at the Individual Level : The Effects of Age on Productivity and Impact. *Journal of the American Society for Information Science and Technology*, 61(8), 1564–1581. doi:10.1002/asi.21348

de Sousa Vieira, E. (2013). *Indicadores bibliométricos de desempenho científico: estudo da aplicação de indicadores na avaliação individual do desempenho científico*. Porto: Universidade do Porto.

Jiménez-Contreras, E., Robinson-García, N., & Cabezas-Clavijo, A. (2011). Productivity and impact of Spanish researchers: reference thresholds within scientific areas. *Revista Española de Documentación Científica*, 34(4): 505-525.

Kissin, I. (2011). A surname-based bibliometric indicator: publications in biomedical journal. *Scientometrics*, 89(1), 273–280. doi:10.1007/s11192-011-0437-3

Kissin, I., & Bradley, E. L. (2013). A surname-based patent-related indicator: the contribution of Jewish inventors to US patents. *Scientometrics*. doi:10.1007/s11192-013-1005-9

Larivière, V. (2012). On the shoulders of the students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, 90(2): 463-481.

Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J., & García, J. A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. *Journal of Informetrics*, 7(2), 412–424. doi:10.1016/j.joi.2013.01.004

van Leeuwen, T.N. (2008). Testing the validity of the Hirsch-index for research assessment purposes. *Research Evaluation*, 17(2): 157-160.

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

Waltman, L., & Eck, N. J. Van. (2013). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 1–23. Retrieved from http://arxiv.org/ftp/arxiv/papers/1203/1203.0532.pdf

Waltman, L., & Van Eck, N.J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2), 406-415. arXiv:1108.3901

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. J. (2010). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37–47. Retrieved from http://arxiv.org/abs/1003.2167

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics*, 87(3), 467–481. doi:10.1007/s11192-011-0354-5

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2011). A Boosted-Trees method for name disambiguation. The Selected Works of Diana Hicks.