Is the Year of First Publication a Good Proxy of Scholars' Academic Age?

Rodrigo Costas[1], Tina Nane[2] and Vincent Larivière[3]

[1] *rcostas@cwts.leidenuniv.nl;* [2] *g.f.nane@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX, Leiden (the Netherlands)

[3] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Station Centre-Ville Montreal, Quebec (Canada)

Abstract
Individual scholars are the central unit of the research system and are increasingly the focus of bibliometric studies. An important aspect in the study of individual scholars is their academic age, which allows for the comparison of scholars that have been academically active in a similar period of time. Based on a sample of Quebec researchers for whom their year of birth, PhD year as well as the year of their first publication are known, we study the relationships among these ages with the aim of determining how their year of first publication can be used to estimate their 'real' age. Moderate correlations have been found among the ages, and the first publication year has a higher correlation with the PhD year than with the birth year. However, an important dispersion of scholars across the different ages is observed; thus, the year of first publication can only be taken as proxy of the real age of scholars. Alternatively, the consideration of cohorts of around 5 years seems to be a reasonable approach. Further research will focus on the exploration of other bibliometric indicators in order to refine the preliminary developments discussed here.

Conference Topic
Methods and techniques

Introduction
In individual-level bibliometric studies, the socio-demographic characteristics of scholars are of central importance to understand and better frame the results obtained (Costas & Bordons, 2011; Gingras, Larivière, Macaluso, & Robitaille, 2008; Mauleón & Bordons, 2006). Among these socio-demographic characteristics we can mention gender (Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013; Mauleón & Bordons, 2006), mobility (Canibano, Otamendy, & Solis, 2011; Franzoni, Scellato, & Stephan, 2012), and nationality (Moed & Halevi, 2014), among others. The development of large-scale author-name disambiguation algorithms (Caron & Van Eck, 2014) as well as the increasing quantity of papers' metadata indexed (e.g. author names and surnames, affiliations, e-mail data, etc.) have allowed the study of the socio-demographic characteristics of scholars at a larger scale. For example, the analysis of the first author names of authors (Larivière et al., 2013) allowed the macro analysis of gender disparities worldwide. The large-scale analysis of the relationship between author names, affiliations and countries collected from scientific publications has open the possibility of studying academic mobility at the world level (Moed, Aisati, & Plume, 2013), as well as the nationality (Costas & Noyons, 2013), migrations (Moed & Halevi, 2014) or even the ethnic origin (Freeman, 2014) of scholars.

A critical element for individual-level bibliometrics is the age of the researchers (Costas & Bordons, 2011; Larivière, Archambault, & Gingras, 2008; Levin & Stephan, 1989), especially from the point of view of its relationship with productivity (Falagas, Ierodiakonou, & Alexiou, 2008; Levin & Stephan, 1989). Age is also a common point of debate in science policy, as it aims to compare scholars of the same 'academic age' (Bornmann & Leydesdorff,

2014). However, one of the main reasons that hinders the development of bibliometric studies at the individual level is the lack of systematic data on the age of scholars, as this information is not systematically collected in bibliographic databases. A commonly used proxy for the study of the age of scholars has been the so-called 'scientific (or academic) age', often defined as the publication year of the first paper of a scholar (Radicchi & Castellano, 2013).[1] The use of this age is very convenient, as it is possible to directly extract it from bibliometric data. However, so far there has not been any analysis on the relationship between this proxy and the real age of scholars. This paper is intended to fill this gap and shed some light on the relationship between the 'bibliometric' age of scholars that can be calculated based on bibliographic information and the 'real' age(s) of individual scholars, namely their birth age and their PhD age. In other words, we aim to infer the birth year and PhD year of scholars based on models that are exclusively based on bibliometric indicators[2] (e.g. first publication year, position of signature, co-authors, etc.). Thus, the main research question can be operationalized as follows: *could the year of first publication (YFP) of a scholar (as recorded in the Web of Science) be considered as a relevant proxy of the birth and/or PhD ages of scholars?*

Methodology

In order to answer the research questions it is necessary to have a dataset of scholars for whom their real ages are certainly known as well as the publication years of their scientific publications. Thus, as our golden set, in this study we have considered one of the (possibly) largest datasets of individual scholars for whom real individual characteristics are known (this dataset has been used in some other studies, e.g. Gingras et al., 2008; Larivière et al., 2011). This dataset is composed by 13,626 university professors from Quebec who have published at least one article during the 1980-2012 period. For every scholar in the dataset, the following individual elements have been codified:

- Year of birth [*Birth year*]
- Year of PhD (year when the scholar has obtained her (first) PhD) [*PhD year*].
- Publication year of their first publication in the Web of Science (WoS) [*YFP*]
- [*Birth year to YFP*], which is calculated as [*YFP*]-[*Birth year*]
- [*PhD year to YFP*] which is calculated as [*YFP*]-[*PhD year*]
- Domain (nine disciplinary fields of activity of the scholar, which is based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP)[3] developed by the U.S. Department of Education's National Center for Education Statistics (NCES).

Complementary, we have also calculated the total number of publications of the scholars in the period 1980-2012 [*P*].

A technical limitation of the dataset is that the WoS publication data starts in 1980, thus meaning that for very old individuals it is not possible to know with certainty if the first publication recorded in the WoS during the period 1980-2012 truly corresponds with their actual first publication. To reduce the effect of this issue, we decided to focus only on those individuals that have a birth year later than 1959 (i.e. we don't expect that many scholars would have a publication before their 20's) and a PhD year also later than 1980 (same criteria

[1] Although this term has also been proposed for the time since the PhD has been awarded (Bar-Ilan, 2014). Some other studies have also focused on the starting year of publication of individuals as proxies of age (Fronczak, Fronczak & Holyst, 2006).

[2] Due to space restrictions, in this paper we focus only on the first publication year as a proxy, and leave for a further version of this paper the consideration of other bibliometric variables.

[3] The Classification of Instructional Programs (CIP) is developed by the U.S. Department of Education's National Center for Education Statistics (NCES). More details can be found at: http://nces.ed.gov/pubs2002/cip2000/

as before). As a result of this filtering we ended up with 3,596 scholars that are the final dataset of our analysis.

Main results

This section presents the main results of the analysis. In Appendix 1 the descriptive scores are presented. Results show that there are differences in individual productivity by domain, which is of course not a surprise. For instance scholars from the Basic Medical Sciences and Health sciences exhibit the highest number of WoS papers, while Humanities the lowest. Similarly, the median birth year of the whole sample is 1965, although there are small differences by domain, with Basic Medical Sciences with the oldest individuals (median=1964) and Social Sciences the youngest (median=1967). The median PhD year of the whole sample is 1998, with the Basic Medical Sciences as the oldest median (1994) and domains such as Business & Management, Education, Non-health professionals getting their PhD on median in 1998.

Regarding the time between the birth of the scholars and the time of their first publication, scholars from Basic Medical Sciences, Engineering, Health Sciences and Science are on median the fastest (32 years) while scholars from Business & Management, Education or Humanities are slower (35 years). From the PhD to the first publication, the fastest are the scholars in Health Sciences (1 year) and the slowest the Humanities (4 years). It is important to keep in mind that here we also have cases with negative values, which means that researchers publish publications before their PhD date; a finding coherent with Larivière (2012).

Relationship between the different ages

In Appendix 2 we present the main correlations between the different ages of the scholars. In Figure 1 a summary of the correlations is presented. In general, there is a reasonably good correlation between birth year and PhD year, and the two real ages of the scholars have moderate correlations with YFP, although the PhD year has a generally better correlation with YFP than the Birth Year. These results suggest that it is reasonable to consider the YFP as a proxy of the scientific age of the researchers.

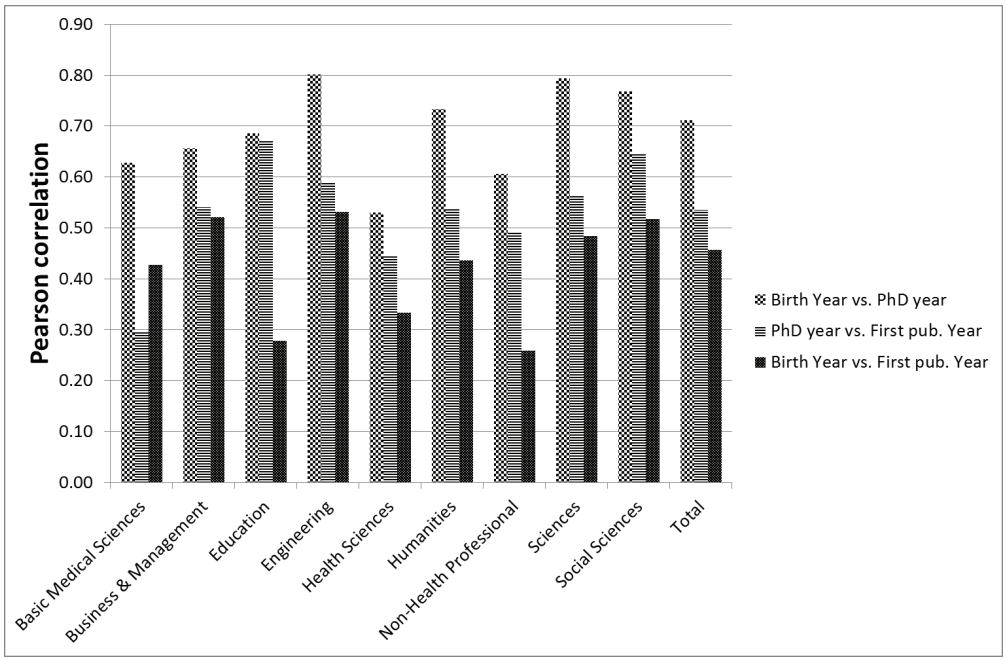

Figure 1. Pearson correlation values of the different ages – by disciplines and all disciplines combined.

YFP as a proxy of the age of researchers

Considering the moderate correlations between the YFP and the real ages of the researchers, we explore the dispersion of the scholars by the different ages. In Figure 2 box plots of each of the three variables (YFP, Birth year and PhD year) grouped by the combination of the same variables are presented. Thus it is possible to understand how scholars distribute across the different ages.

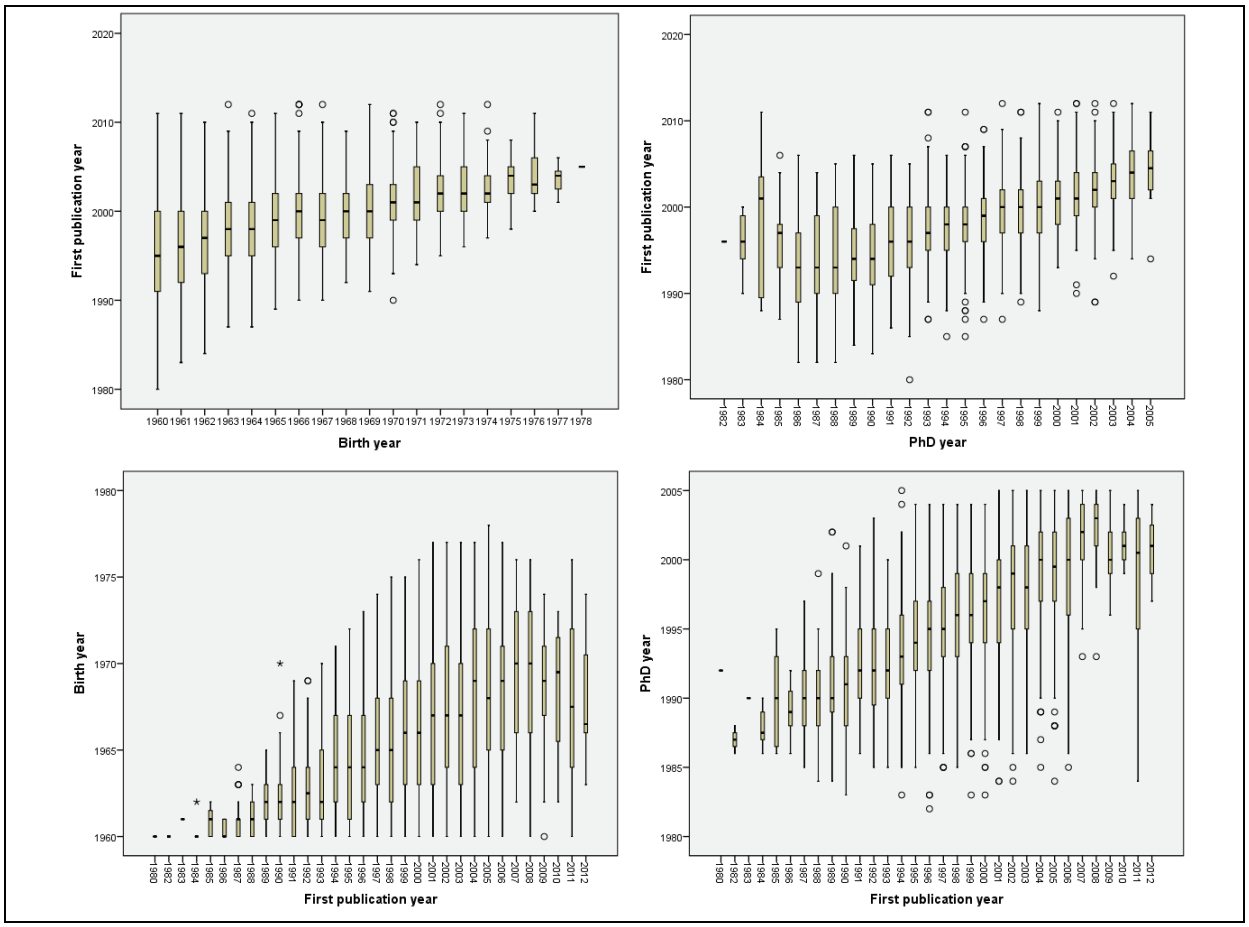

Figure 2. Box plot distribution of scholars across the different ages (all scholars together).

The two graphs on top of Figure 2 present boxplots of YFP observations grouped by each distinct birth year and PhD year. In the case of the birth year, it is possible to see how the earlier the year of birth the larger the variation of the YFP, thus indicating how researchers of all ages start their publication activities at different points in their lives, although the majority (i.e. the 'box' in the graph) tends to concentrate in a range of 5 to 10 years. The YFP median also tends to increase as the birth year increases. In the case of the PhD year we see also a quite disperse distribution of the first publication year of the scholars, although (with the exception of some irregularities among the scholars with the earliest PhD years) we notice a stepper increase in the median value of the YFP as the PhD year increases.

The graphs on the bottom of Figure 2 show the distribution of the two real ages (birth and PhD years) as a function of the YFP. Here we can also see an important dispersion of scholars across the two ages. However, in order to summarize the results of these two graphs, in Figure 3 we present the interquartile ranges (i.e. range of the number of years that include the 50% of all the observations), thus allowing to identify where most of the scholars are located in the distribution as a function of their first publication year.

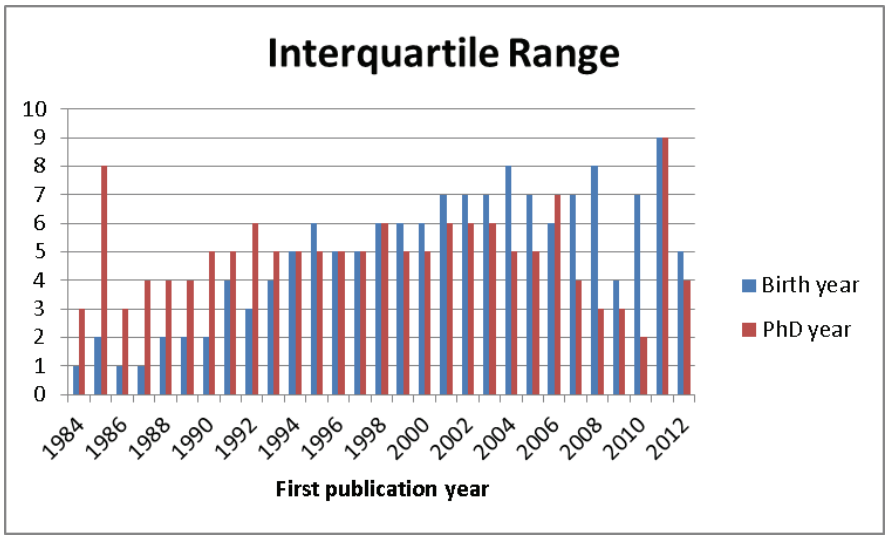

Figure 3. Interquartile range (in number of years) for Birth year and PhD year as a function of YFP.

Figure 3 shows that the interquartile range in all cases is smaller than 10 years for any of the two ages considered. Actually the average for all the YFP years considered is 4.9 years for both ages (with a median of 5). Thus, a possible interpretation of this result is that if we would only count with the YFP of the scholars, with a range of around 5-10 years we would be able to capture the real age of about 50% of all the scholars who started to publish that year.

Exploring a predictive model for the age of scholars based on bibliometric indicators

In this section a more predictive approach is presented. We are interested in estimating the birth and PhD years of a generic researcher by using the YFP indicator in our data sample. Numerous approaches can be taken, from the selection of different models and independent variables that could influence the two ages. In the present study we choose the simple linear regression model, with the average birth year and the average PhD year as dependent variables and the YFP as the independent variable. We will therefore infer on the average birth and PhD year of a scholar, and Figures 4 and 5 provide the linear regression fit of the two models, along with confidence and prediction intervals.

Using linear regression analysis the average ages (birth year and PhD year) of the whole list of scholars are fitted, including a 95% confidence interval as well as a 95% prediction interval. Although both intervals account for the uncertainty of the regression parameter estimates, there is an important distinction between the two intervals. The confidence interval is supposed to cover the true average birth year (of all the scholars in the statistical population) with high probability in 95% of the cases. The prediction interval provides limits on a future sampled observation that is an average of a given number of scholars from the set of all the scholars in the world. The prediction intervals refer then to actual observations in the data, and hence account also for the variation in the data, whereas the confidence intervals refer to the population's (of all scholars) average birth year. The prediction intervals are always larger than the confidence intervals.

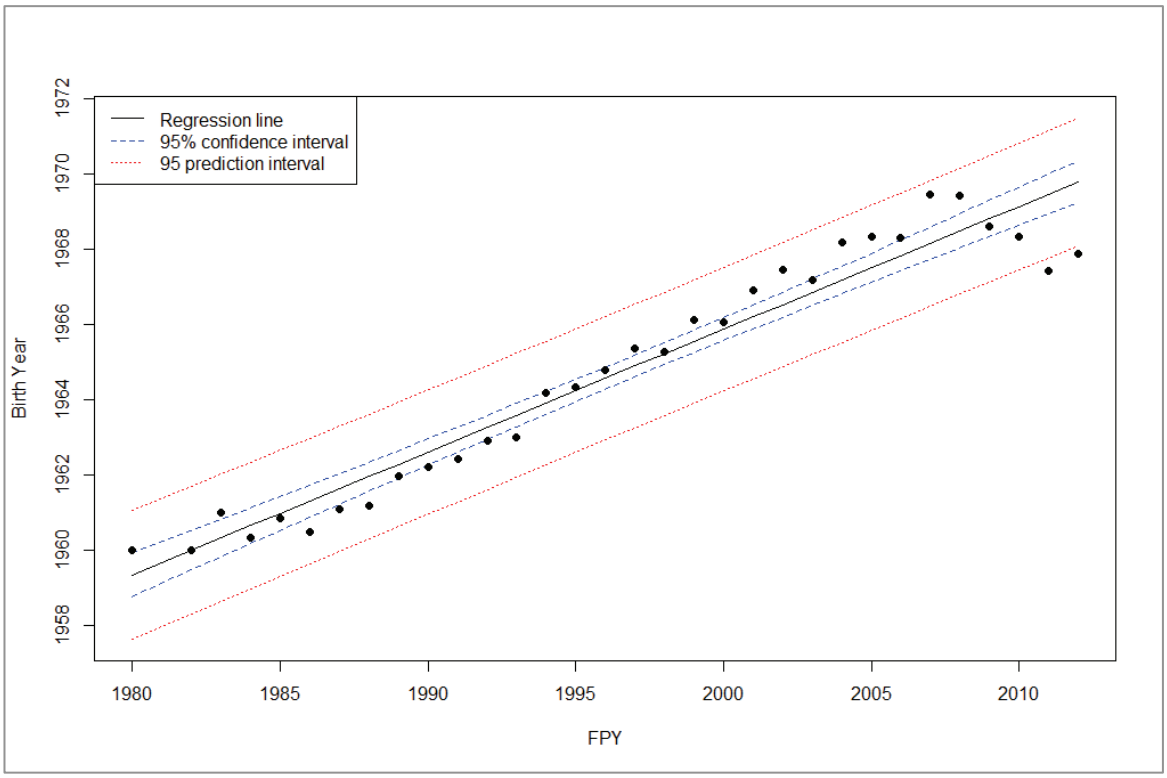

Figure 4. Average birth year by YFP, fitting a regression line and 95% confidence and prediction intervals.

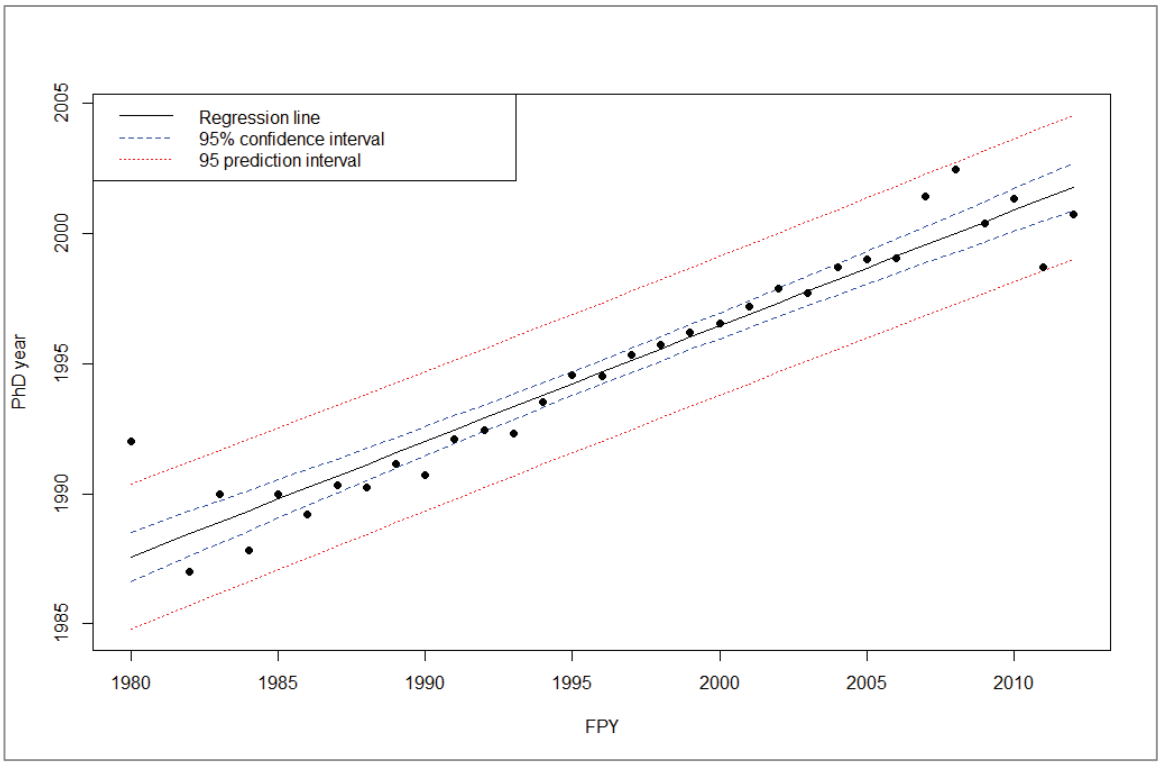

Figure 5. Average PhD year by YFP, fitting a regression line and 95% confidence and prediction intervals.

The main difference with bottom graphs in Figure 2 is that here the target is to estimate the average age of scholars from a given YFP. For example, in Figure 4 we can see how for scholars with a YFP=1995 their average birth year would be 1963, and the prediction interval ranges between 1961 and 1965. A similar pattern is observed in Figure 5 for PhD year (i.e.

with YFP=1995 the average PhD year would range around a period of five years). This suggests that we would be able to estimate the average ages of the scholars with a given YFP within an interval of 5 years. Of course, it is important to keep in mind that this analysis is based on the average values for all scholars, which is different from the individual prediction of individual scholars; however the relatively short prediction intervals (around 5 years) supports the importance of the YFP as relevant proxy for the ages of individual scholars.

Discussion and conclusions

Age is one of the most important socio-demographic determinants of researchers' activities, funding, output and impact. However, the lack of systematically recorded information on the age (real or academic) of researchers makes the need of developing reliable and valid proxies a priority. So far, the age of the first publication of individual scholars has been frequently considered as a proxy of the real age of scholars; however its validity has never been tested. Based on a sample of Quebec researchers for whom their actual birth year, PhD year as well as the year of their first publication are known, a study on the relationships among these ages has been performed.

The three ages correlate moderately well, birth year and PhD year have a good relationship, and YFP has moderate correlations with the other two ages, particularly with the PhD year. It is also possible to detect an important dispersion of scholars across the different ages, indicating that new authors (and new researchers) basically can come from a wide range of years. This means that, in spite of the moderate correlation between the YFP and the other ages, the YFP can only be considered as a proxy for researchers' age, as it does mix researchers with different birth and PhD years. The consideration of cohorts of years seems to be a more reasonable alternative. Thus, it is possible to argue that considering authors who started to publish in a given year, the majority of these scholars would have ages (birth and PhD) within a range of 5 to 10 years.

It is important also to highlight some of the limitations of this study. In the first place, we are working with a dataset of scholars from only one location (Quebec in Canada), so we need to keep in mind the limitations of the representativeness of our sample for the whole world. Thus, issues related with the changes and internal evolution of PhD programs could partly influence the results and hinder their generalization. Secondly, WoS is the only database considered for the determination of the YFP, however scholars can publish outputs not covered by this database, which is likely the case in Quebec, whose local literature in the social sciences and humanities is highly relevant (Larivière & Macaluso, 2011). Thirdly, in this study we haven't explored differences across fields, but arguably there are differences in the relationship between the ages and the first publication year of the scholars as disciplinary differences in individual productivity have been also discussed (Ruiz-Castillo & Costas, 2014).

All in all, considering the limitations previously exposed, our results are still policy-relevant and support the idea that the first publication year(s) of individual scholars can work as a reasonable proxy as their age, particularly when considering cohorts of researchers. For the final version of the paper other approaches will be also considered, including the analysis of the positions of the scholars in the papers (as these positions are related with the age of scholars (Costas & Bordons, 2011), other bibliometric indicators (e.g. the total number of publications of a scholar and total number of citations, which are age dependent) as well as the different disciplines of scholars. Finally, the consideration of other datasets from other countries and/or disciplines is an important development in order to globally validate the different tests and models obtained and to establish a more generalizable approach for the estimation of ages based on bibliometric data. A potential recommendation derived from this study is the relevance of incorporating information about the age, PhD year, gender and other

demographic characteristics in modern Research Information Systems (RIS). This would allow for more accurate studies of the demographics and changes in the trends of scientific productivity of individual scholars.

References

Bar-Ilan, J. (2014). Evaluating the individual researcher - adding an altmetric perspective. *Research Trends*, *37*, 31–34.

Bornmann, L. & Leydesdorff, L. (2014). On the meaningful and non-meaningful use of reference sets in bibliometrics. *Journal of Informetrics*, *8*(1), 273–275. doi:10.1016/j.joi.2013.12.006

Canibano, C., Otamendy, F. J. & Solis, F. (2011). International temporary mobility of researchers: a cross-discipline study. *Scientometrics*, *89*(2), 653–675.

Caron, E. & Van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *19th International Conference on Science and Technology Indicators. "Context counts: pathways to master big data and little data."* Leiden: CWTS-Leiden University.

Costas, R. & Bordons, M. (2011). Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. *Scientometrics*, *88*(1), 145–161. Retrieved from http://www.springerlink.com/index/10.1007/s11192-011-0368-z

Costas, R. & Noyons, E. (2013). Detection of different types of " talented " researchers in the Life Sciences through bibliometric indicators : methodological outline Sciences through bibliometric indicators : methodological outline 1. *CWTS Working Paper Series*, (CWTS-WP-2013-006). Retrieved from http://www.cwts.nl/pdf/CWTS-WP-2013-006.pdf

Falagas, M. E., Ierodiakonou, V. & Alexiou, V. G. (2008). At what age do biomedical scientists do their best work? *The FASEB Journal*, *22*(12), 4067–4070.

Franzoni, C., Scellato, G. & Stephan, P. (2012). Patterns of international mobility of researchers : evidence from the GlobSci survey. In *International Schumpeter Society Conference* (pp. 1–32). Retrieved from http://www.aomevents.com/media/files/ISS 2012/ISS SESSION 7/Scellato.pdf

Freeman, R. B. (2014). Strength in diversity. *Nature*, *513*, 305.

Fronczak, P., Fronczak, A. & Holyst, J. A. (2006). Publish or perish: analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. *arXiv*.

Gingras, Y., Larivière, V., Macaluso, B. B., Robitaille, J.-P. & Lariviere, V. (2008). The Effects of aging on researchers' publication and citation patterns. *Plos ONE*, *3*(12), e4048. doi:10.1371/journal.pone.0004048

Lariviere, V., Archambault, E. & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: from exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288–296. doi:10.1002/asi

Larivière, V., Gingras, Y., Cronin, B. & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*, 4–6.

Levin, S. G. & Stephan, P. E. (1989). Age and research productivity of academic scientists. *Research in Higher Education*, *30*(5), 531–549.

Mauleón, E. & Bordons, M. (2006). Productivity , impact and publication habits by gender. *Scientometrics*, *66*(1), 199–218.

Moed, H. F., Aisati, M. M. & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*, 929–942. doi:10.1007/s11192-012-0783-9

Moed, H. F. & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 1987–2001. doi:10.1007/s11192-014-1307-6

Radicchi, F. & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, *97*(3), 627–637. doi:10.1007/s11192-013-1027-3

Ruiz-Castillo, J. & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*(4), 917–934. doi:10.1016/j.joi.2014.09.006

Appendix 1. Main descriptive values

Disciplinary division		Birth year	PhD year	YFP	P	Birth year to YFP	PhD year to YFP
Basic Medical Sciences	N	713	713	713	713	713	713
	Mean	1993.66	1964.72	1997.01	52.54	32.29	3.34
	Std. Deviation	4.503	3.427	4.835	67.27	4.58	5.54
	Median	1994.00	1964.00	1997.00	30.00	32	3
	Minimum	1983	1960	1980	1	20	-13
	Maximum	2005	1976	2008	788	46	21
Business & Management	N	243	243	243	243	243	243
	Mean	1997.50	1965.92	2000.56	10.92	34.64	3.05
	Std. Deviation	4.427	4.313	4.476	12.405	4.31	4.269
	Median	1998.00	1965.00	2001.00	7.00	35	3
	Minimum	1983	1960	1986	1	25	-10
	Maximum	2005	1976	2012	96	49	27
Education	N	47	47	47	47	47	47
	Mean	1997.38	1965.49	2001.04	8.43	35.55	3.66
	Std. Deviation	3.943	4.117	5.254	13.333	5.71	3.93
	Median	1998.00	1965.00	2001.00	4.00	35	3
	Minimum	1989	1960	1986	1	25	-5
	Maximum	2003	1974	2010	70	48	12
Engineering	N	514	514	514	514	514	514
	Mean	1996.38	1966.27	1998.67	38.08	32.40	2.30
	Std. Deviation	4.713	4.488	4.509	48.889	4.36	4.19
	Median	1996.00	1966.00	2000.00	24.50	32	2
	Minimum	1982	1960	1985	1	22	-11
	Maximum	2005	1977	2009	692	44	17
Health Sciences	N	292	292	292	292	292	292
	Mean	1996.89	1965.45	1998.10	49.80	32.65	1.20
	Std. Deviation	4.183	4.006	4.800	72.488	5.13	4.76
	Median	1997	1965	1998	30	32	1
	Minimum	1985	1960	1984	1	22	-13
	Maximum	2005	1976	2012	788	49	18
Humanities	N	347	347	347	347	347	347
	Mean	1996.78	1965.76	2001.11	3.91	35.35	4.32
	Std. Deviation	4.341	4.115	4.382	5.338	4.52	4.19
	Median	1997	1965	2001	2	35	4
	Minimum	1986	1960	1986	1	24	-6
	Maximum	2005	1978	2012	65	47	20
Non-Health Professional	N	112	112	112	112	112	112
	Mean	1997.84	1965.52	2001.21	10.30	35.70	3.36
	Std. Deviation	4.594	4.480	5.070	14.222	5.84	4.89
	Median	1998	1965	2001.5	4	35	3

Disciplinary division		Birth year	PhD year	YFP	P	Birth year to YFP	PhD year to YFP
	Minimum	1985	1960	1990	1	24	-6
	Maximum	2005	1977	2012	70	51	21
Sciences	N	826	826	826	826	826	826
	Mean	1995.35	1965.88	1997.92	36.45	32.04	2.57
	Std. Deviation	4.441	4.287	4.860	48.406	4.67	4.37
	Median	1996	1965	1999	25.00	32	3
	Minimum	1985	1960	1982	1	22	-11
	Maximum	2005	1977	2012	775	46	17
Social Sciences	N	502	502	502	502	502	502
	Mean	1997.36	1966.75	1999.66	15.87	32.9084	2.3008
	Std. Deviation	4.25	4.33	4.53	19.11	4.36	3.7
	Median	1998.00	1967.00	2000.00	10.00	33	2
	Minimum	1987	1960	1986	1	23	-11
	Maximum	2005	1977	2012	204	48	15
Total	N	3596	3596	3596	3596	3596	3596
	Mean	1995.95	1965.77	1998.73	32.04	32.97	2.78
	Std. Deviation	4.64	4.18	4.89	50.56	4.77	4.60
	Median	1996	1965	1999	17	33	3
	Minimum	1982	1960	1980	1	20.00	-13.00
	Maximum	2005	1978	2012	788	51.00	27.00

Appendix 2. Pearson correlations by ages

Division	Ages	Birth year	YFP	PhD year
Basic Medical Sciences	Birth year	1.000	0.426	0.627
	First publication year	0.426	1.000	0.297
	PhD year	0.627	0.297	1.000
Business & Management	Birth year	1.000	0.521	0.656
	First publication year	0.521	1.000	0.540
	PhD year	0.656	0.540	1.000
Education	Birth year	1.000	0.277	0.686
	First publication year	0.277	1.000	0.670
	PhD year	0.686	0.670	1.000
Engineering	Birth year	1.000	0.531	0.800
	First publication year	0.531	1.000	0.588
	PhD year	0.800	0.588	1.000
Health Sciences	Birth year	1.000	0.333	0.530
	First publication year	0.333	1.000	0.444
	PhD year	0.530	0.444	1.000
Humanities	Birth year	1.000	0.435	0.733
	First publication year	0.435	1.000	0.538
	PhD year	0.733	0.538	1.000
Non-Health Professional	Birth year	1.000	0.258	0.605
	First publication year	0.258	1.000	0.492
	PhD year	0.605	0.492	1.000
Sciences	Birth year	1.000	0.484	0.793
	First publication year	0.484	1.000	0.561
	PhD year	0.793	0.561	1.000
Social Sciences	Birth year	1.000	0.517	0.768
	First publication year	0.517	1.000	0.646
	PhD year	0.768	0.646	1.000
Total	Birth year	1.000	0.457	0.711
	First publication year	0.457	1.000	0.535
	PhD year	0.711	0.535	1.000