Qualitative and Quantitative Methods in Libraries (QQML) 11,1:133-148, 2022

Understanding the h-index (Hirsch Index) and its Correlation with the Number of Citations and Number of Documents

Amit Kumar[1] & Abhay Maurya[2]

[1 2]Department of Library and Information Science, Mizoram University, Aizawl

Abstract: The impact of the work of any researcher is vital for a variety of reasons. While it helps the universities to decide the appointment of researchers in the position of faculty, it also helps the funding organizations to decide on the funding of their research works. Further, the impact helps the award-giving institutions, like the Nobel Committee to decide whether or not to confer awards to the researchers. Though a variety of indexes help the organizations and students of scientometry to gauge the impact of any researcher, the Hirsch index (h-index) is the most popular of the various metrics. The h-index is calculated based on the number of citations the scientific productions of the researchers have received from other future researchers as also the number of scientific documents produced by them. With the dataset comprising of Nobel Laureates in Chemistry from 2014 till 2019, this study attempts to correlate the number of citations and the number of documents with the h-index. The level of correlation of the independent variables with the h-index has also been assessed as a part of this study. This study has observed that both the number of citations and the number of scientific productivity have a direct correlation with the h-index though the number of citations is a better fit. This study has also observed the validity of Yong's formula.

Keywords: Hirsch Index, Nobel Laureates, Impact, Researcher, Citations, Production, Correlation

Received: 22.2.2022 Accepted: 1.4.2022 ISSN 2241-1925

1. Introduction

In 2005, J E Hirsch proposed a measure that quantifies the impact of any scientist[1]Normally called the h-index, this index finds references in various citation databases which include Scopus, Web of Science, Google Scholar, and the like. This index has helped in the progress of science by helping the scientific community by way of hiring, making funding decisions, and promotion (Abbott et. al., 2010; McNutt, 2014; Hicks et. al., 2015). Since the time the h-index has been proposed, it has undergone several modifications. Besides scholars have also proposed several alternatives (Panaretos and Malesios, 2009; Sinatra et. al, 2016), but none could replace the h-index from being used to measure the publication output of any scientist. Several factors can be held responsible for the survival of the h-index as a measure of the output. While the h-index provides the summary of the output of any scientist using numerical values that make both ranking and comparison easier, the h-index of any scientist can be calculated at any time of his/her career. The ease of calculation using the number of citations, and the ease of interpretation has prompted scholars to continue using the h-index. Though several scholars have criticized the h-index, it is still one of the most widely used indexes to measure the productivity of any scientist (Hirsch,2007; Radicchi et. al., 2008; Henzinger et. al., 2009; Acuna et. al., 2012).

According to the definition provided by J E Hirsch during the process of proposing the index, the h-index depends upon the number of scientific publications produced by any scientist and the number of citations received by these publications. Using the data of Nobel Laureates in Chemistry from 2014 till 2019, we have tried to assess the dependence of the h-index on the number of publications and the number of citations. The relevant data for the analysis has been extracted from Scopus.

2. Literature Review

Sinatra et al. (2016) used nearly one million scientific publications produced by 2887 physicists present in the American Physical Society dataset and 7630 scientists in the Web of Science database to compare and rate them. The study was an attempt to correlate the impact indicators with scientific awards. However, the study suffers from the limitation that it was restricted to the scientific productivity of a few scientists. Further, the study used the Nobel Prize in Physics and the Dirac and Boltzmann Medals as the indicators that portray scientific repute. A similar study has also been conducted by Ioannidis

et al. (2016)when he analysed both the citation and publication data of 84116 scientists to see if they compare with the Nobel Laureates from 2011 till 2015. Similarly, Ayaz and Masood (2020) evaluated the research impact of 236416 publications in the field of computer science. The authors based their study on the Scientometric indexes of 47 awardees in the list. Koltun & Hafner (2021) analysed highly cited scientific articles written by 4000 different researchers from four separate domains. The selected researchers have won more than 1500 different awards. The authors observed that the effectiveness of Scientometric indexes has been declining over the years. The study cited that the correlation of the h-index with awards in physics including the prestigious Nobel Prize has decreased from 0.34 in 2010 to 0 in 2019. The reason for this decline can be attributed to the non-adherence of fractional citation among the co-authors. O. Wilke (2014) tested the formula proposed by Yong to calculate the h-index with the total number of citations on 29 researchers attached to the Department of Integrative Biology and observed that the proposed formula has worked fairly well. He extracted both the total citations and the h-index of individual researchers from Google Scholar and has covered a wide range of career stages. The calculation has been visualized in the form of a graph.

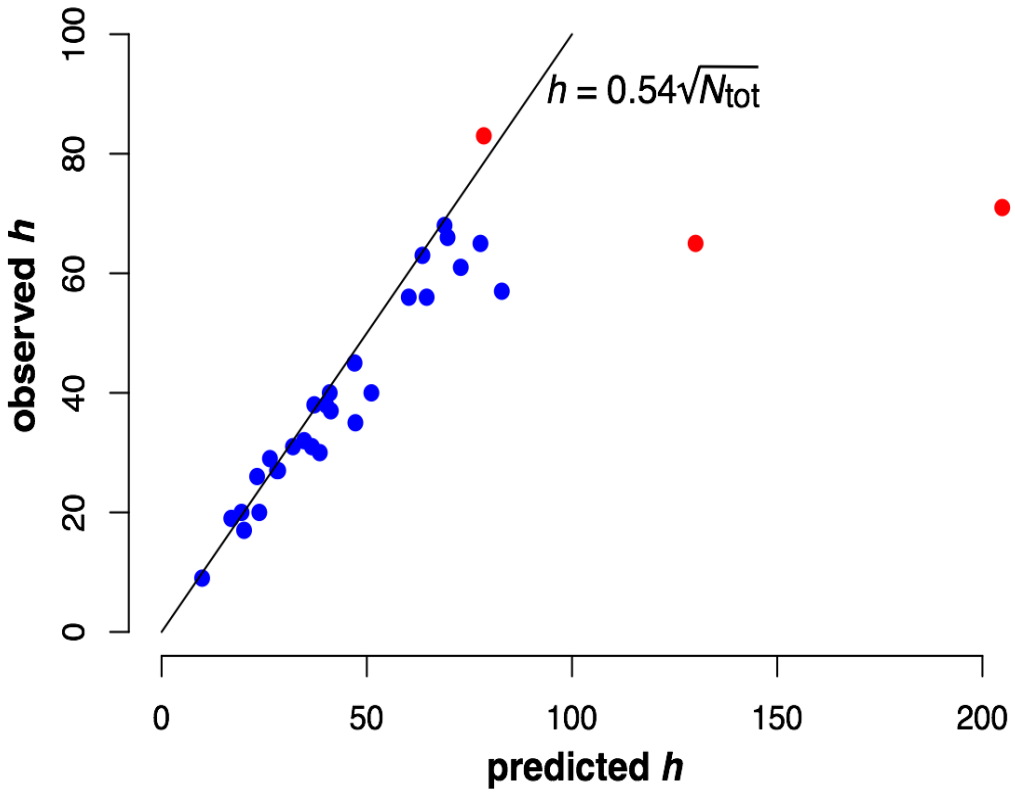

Fig 1: Observed vs predicted h-index (Source: O. Wilke[11])

The dots in red belong to the highly cited faculty of the department. In another study, Mahmoudi et al. (2021) proposed a statistical model to calculate the h-index based upon the number of citations and the number of years that have elapsed since the publication of the first paper using simple non-linear regression. The study observed that while both the number of citations and the time plays a significant role in the value of the h-index, the number of citations was a better fit than the number of years that have elapsed since the publication of the first paper.

3. Aim of The Study

With the dataset comprising the h-index, the number of scientific publications, and the number of times these publications have been cited by future researchers for Nobel Laureates in Chemistry for the years 2014 till 2019, this study aims to understand the correlation between the independent variables

(the number of scientific publications, and the number of times these publications) with the dependant variable (h-index).

4. Motivation Behind The Study

In his paper titled Critique of Hirsch's Citation Index: A Combinatorial Fermi Problem, Alexander Yong (2014) has derived a mathematical formula that correlates the h-index with the total citations received by the publications produced by the scientist. The paper argues that for any scientist the h-index can be calculated using the formula

$$h - index = 0.54\ X\ \sqrt{N_{Total}}$$

where N_{Total} represents the total number of citations.

Yong has estimated the h-index and has shown that scientists producing only a few highly cited papers have values of h-index below the estimate. Though various scholars have criticized the h–index as being biased towards the highly cited articles, Yong observed that although the h-index measures the total amount of citations, it does not depend upon the highly cited publications. The motivation behind this study lies in applying the results obtained by Yong to Nobel Laureates and also to analyze the dependence of the h-index on the total number of publications.

5. Methodology

The relevant data which includes the number of scientific production and citations have been extracted from https://www.scopus.com and all calculations and visualizations have been done using Microsoft excel.

6. Results

This study has considered all Nobel Laureates in Chemistry from 2014 till 2019. As per the will left behind by Alfred Nobel, a maximum of three researchers can be awarded the Nobel Prize in any domain during any year. In consonance with this covenant, three researchers have been awarded the Nobel

Prize during the period under study. Table 1 provides the data of all the Nobel Laureates considered as a part of this study.

Table 1: h-Index, number of documents, and citations

YEAR	NAME	h-INDEX	DOCUMENTS	CITATIONS
	ERIC BETZIG	61	137	25483
2014	STEFAN WALTER HELL	99	445	46777
	WILLIAM ESCO MOERNER	78	452	29860
	TOMAS ROBERT LINDAHL	99	238	38916
2015	PAUL LAWRENCE MODRICH	79	188	22816
	AZIZ SANCAR	107	414	38982
	JEAN-PIERRE SAUVAGE	99	505	37393
2016	SIR JAMES FRASER STODDART	139	1087	103411
	BERNARD LUCAS FERINGA	120	855	63335
	JACQUES DUBOCHET	56	143	12557
2017	JOACHIM FRANK	91	388	30131
	RICHARD HENDERSON	31	148	21336
2018	FRANCES HAMILTON ARNOLD	101	349	37177
	GEORGE PEARSON SMITH	29	55	5412
	SIR GREGORY PAUL WINTER	86	203	32293
2019	JOHN BANNISTER GOODENOUGH	144	968	111407
	MICHAEL STANELY WHITTINGHAM	69	336	26712
	AKIRO YOSHINO	12	69	1284

It can be observed that John Bannister Goodenough who was awarded the Nobel Prize in Chemistry in 2019 has the highest h-index of 144 followed by Sir James Fraser Stoddart a Nobel Laureate of 2016 who had an h-index of 139. While Bernad Lucas Feringa, another Nobel Laureate of 2016 has an h-index of 120, Aziz Sancar, who was awarded the Nobel Prize in 2015 has an h-index of 107. Among the number of documents, Table 1 shows that Sir James Fraser Stoddart has produced the highest number of documents at 1087 followed by John Bannister Goodenough who has produced 968 scientific documents. The data regarding the number of citations is also no different. While the scientific productions of John Bannister Goodenough received the highest number of citations at 111407, the scientific productions of Sir James Fraser Stoddart have received 103411 citations.

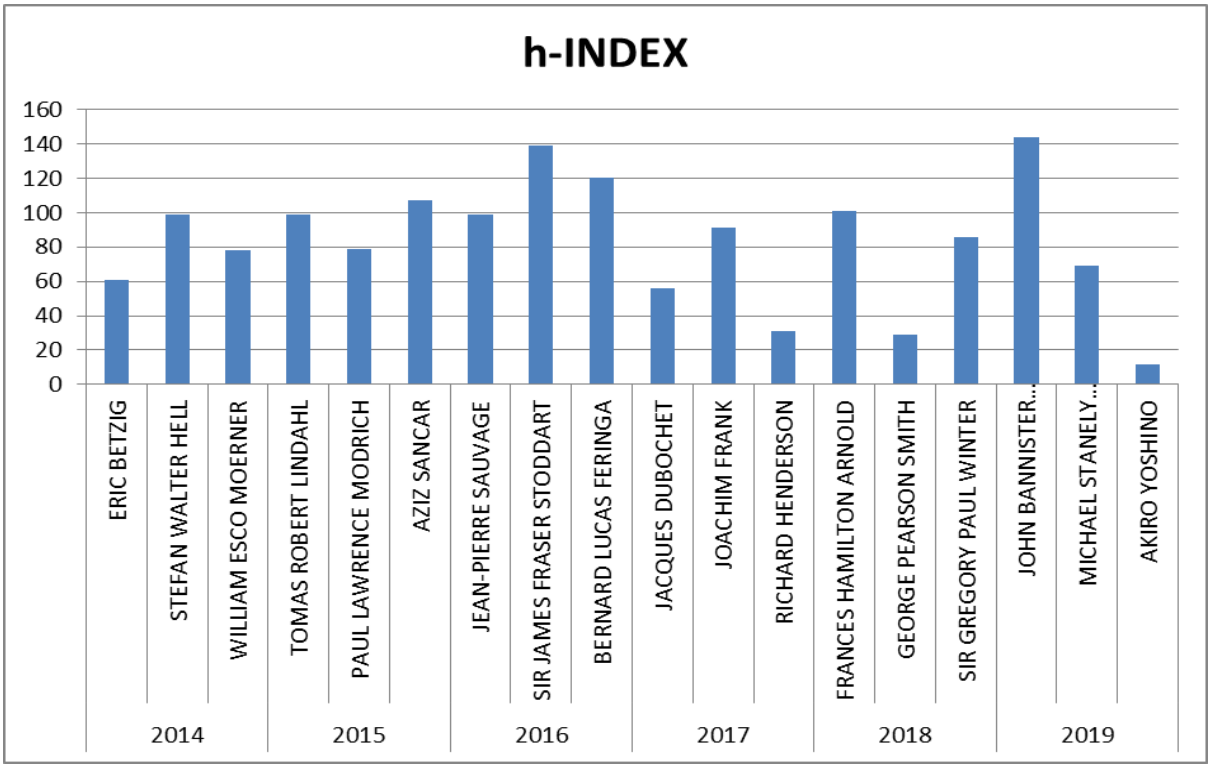

Fig 2: Graph showing h-index of Chemistry Nobel Laureates

Figure 2 is a graphical representation of the h-indexes of the Nobel Laureates. An analysis of the graph shows that John Bannister Goodenough has the highest h-index among the Nobel Laureates considered for the study followed by Sir James Fraser Stoddart.

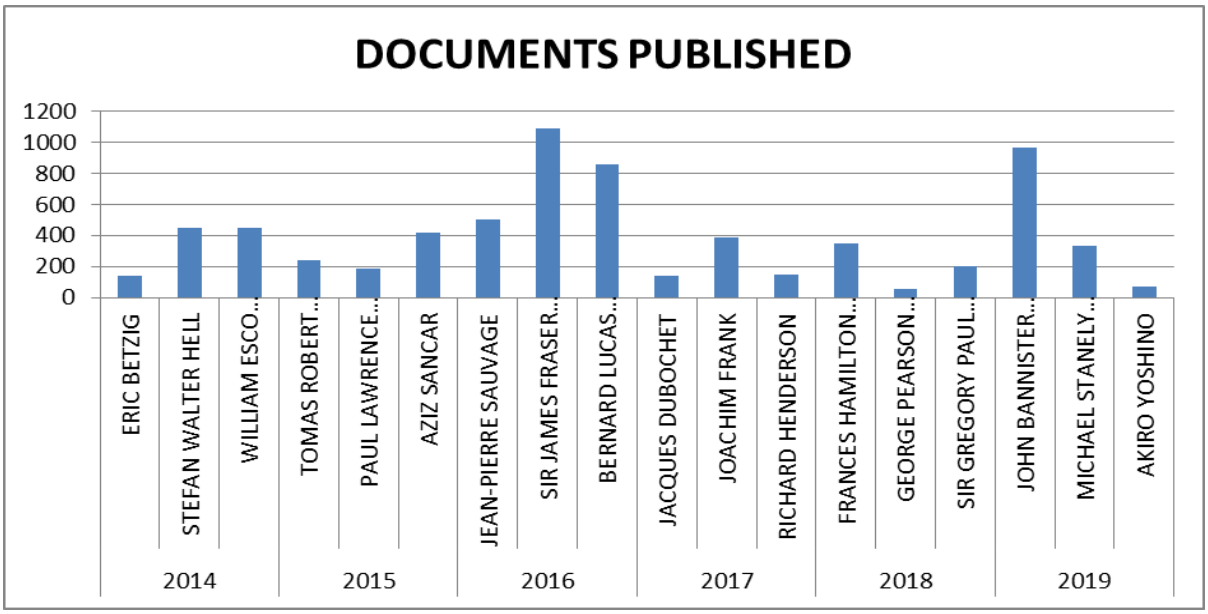

Fig 3: Graph showing the number of documents published by Chemistry Nobel Laureates

A look into the number of documents published by the Nobel Laureates also shows the same pattern. While Sir James Stoddart Fraser has published the highest number of documents, the number of documents published by John Bannister Goodenough and Bernard Lucas Feringa is not far behind. A similar trend is also observed as regards the number of citations received by the Nobel Laureates.

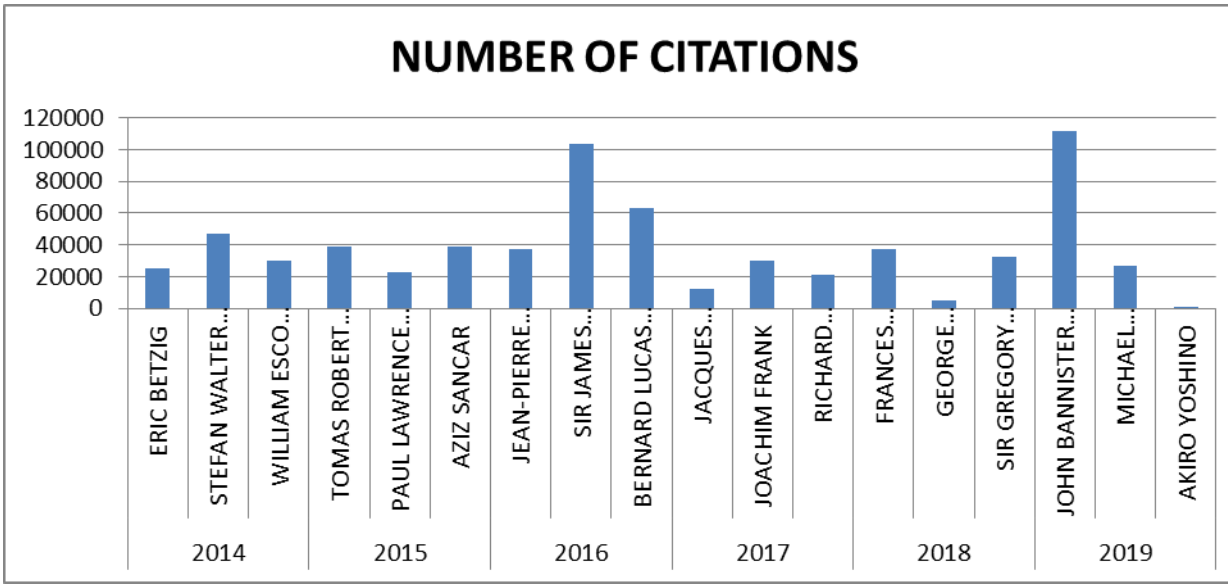

Fig 4: Graph showing the number of citations received by Chemistry Nobel Laureates

A look into the graph depicting the correlation between the h-index and the number of documents published by the Nobel Laureates bears testimony to the fact that a higher number of publications result in a higher value of the h-index.

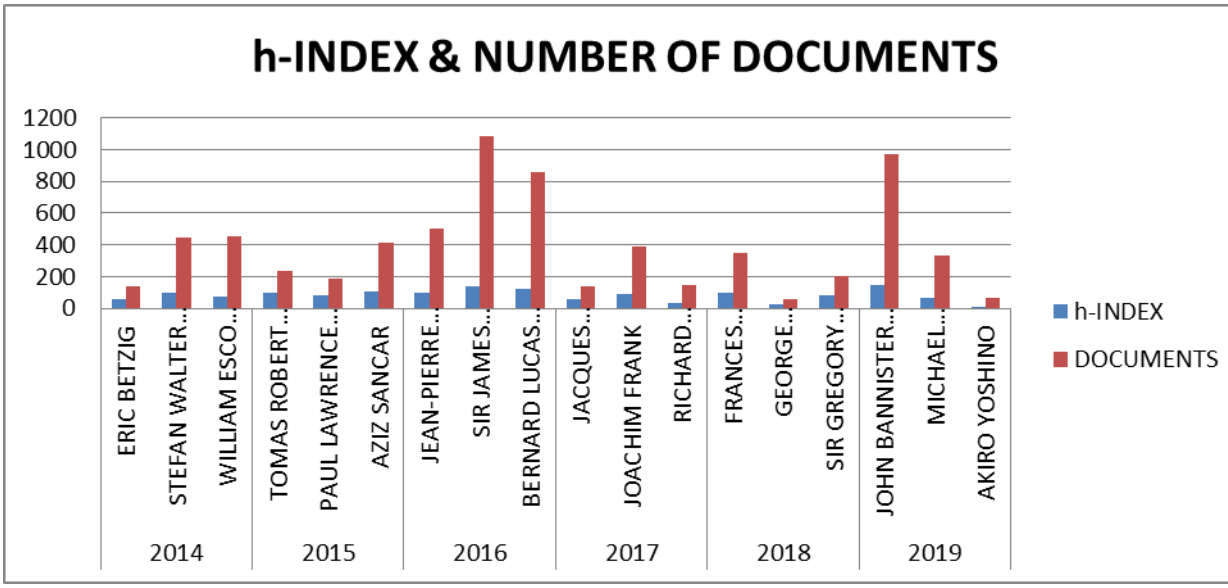

Fig 5: Graph showing the h-index and number of documents published by Chemistry Nobel Laureates

Analysis of the graph showing the h-index and the number of citations also show that the h-index bears a direct correlation with the number of citations, which is following the observation made by Alexander Yong.

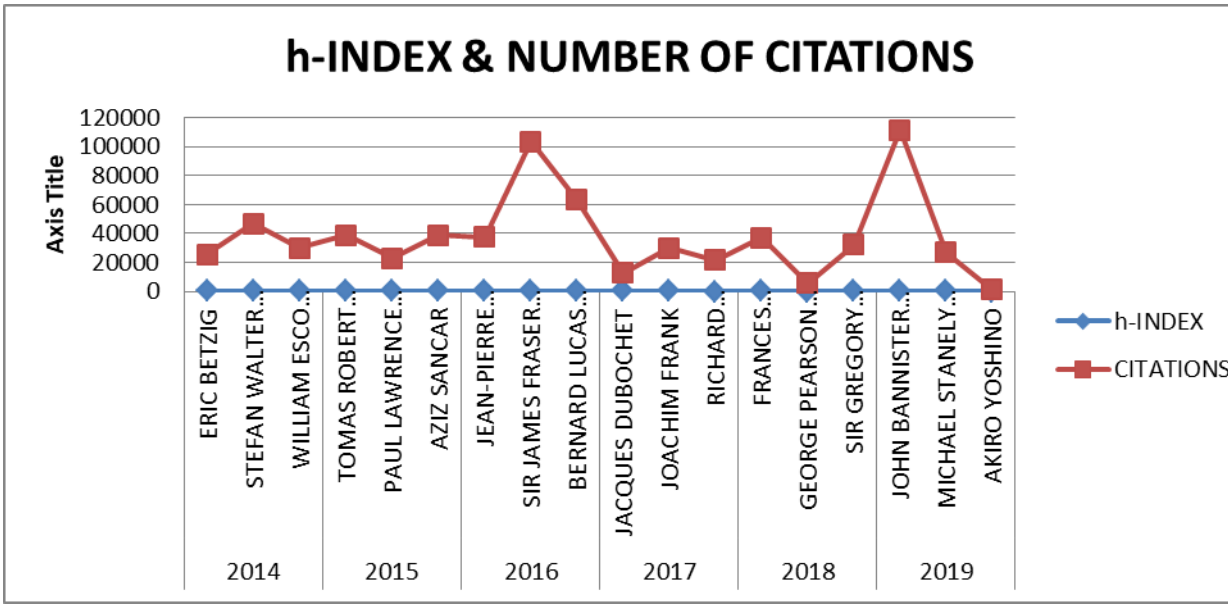

Fig 6: Graph showing the h-index and number of citations received by Chemistry Nobel Laureates

Figure 7 shows how the number of documents published by any researcher and the number of citations received by those publications is correlated with the h-index.

	Citation1	Citation2	Citation3	Citation4	Citation5	Citation6	Citation7	Citation8	Citation9	Citation10	Citation11	Citation12	Citation13	Citation14	Citation15	Citation16	Citation17	Citation18
Document 1	1																	
Document 2	0.9999926	1																
Document 3	0.9999651	0.9999899	1															
Document 4	0.9999999	0.9999944	0.9999693	1														
Document 5	0.9999988	0.9999974	0.999977	0.9999954	1													
Document 6	0.9999909	0.9999999	0.9999917	0.9999929	0.9999963	1												
Document 7	0.9999773	0.9999958	0.9999887	0.9999806	0.9999866	0.999997	1											
Document 8	0.9999868	0.9999992	0.9999948	0.9999833	0.9999936	0.9999996	0.9999987	1										
Document 9	0.9999716	0.9999932	0.9999997	0.9999754	0.9999822	0.9999947	0.9999997	0.9999971	1									
Document 10	0.999994	0.9999999	0.999988	0.9999957	0.9999982	0.9999997	0.9999946	0.9999986	0.9999917	1								
Document 11	0.9999819	0.9999977	0.9999973	0.9999849	0.9999901	0.9999985	0.9999997	0.9999996	0.9999988	0.9999967	1							
Document 12	0.9999976	0.9999986	0.9999809	0.9999986	0.9999998	0.9999978	0.9999896	0.9999956	0.9999856	0.9999992	0.9999926	1						
Document 13	0.9999948	0.9999998	0.9999868	0.9999963	0.9999986	0.9999994	0.9999938	0.9999982	0.9999907	1	0.9999961	0.9999995	1					
Document 14	0.9999987	0.9999975	0.9999772	0.9999994	1	0.9999964	0.9999868	0.9999937	0.9999824	0.9999983	0.9999903	0.9999998	0.9999987	1				
Document 15	0.9999998	0.9999946	0.9999696	1	0.9999995	0.9999931	0.9999809	0.9999895	0.9999757	0.9999958	0.9999851	0.9999987	0.9999965	0.9999995	1			
Document 16	0.9999926	1	0.9999898	0.9999945	0.9999974	0.9999999	0.9999958	0.9999992	0.9999932	0.9999999	0.9999977	0.9999986	0.9999998	0.9999975	0.9999946	1		
Document 17	0.9999812	0.9999974	0.9999975	0.9999842	0.9999896	0.9999983	0.9999998	0.9999995	0.999999	0.9999964	1	0.9999922	0.9999958	0.9999897	0.9999845	0.9999974	1	
Document 18	0.9993125	0.9994478	0.9995872	0.9993314	0.9993694	0.9994618	0.9995397	0.9994898	0.9995635	0.9994346	0.9995173	0.9993908	0.9994267	0.9993705	0.9993331	0.9994476	0.999521	1

Fig 7: Figure showing the correlation between h-index, number of documents, and citations

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.851729109							
R Square	0.725442474							
Adjusted R Square	0.708282629							
Standard Error	19.46331796							
Observations	18							
ANOVA								
	df	SS	MS	Significance F				
Regression	1	16014.86806	16014.86806	42.27558345	7.29665E-06			
Residual	16	6061.131937	378.820746					
Total	17	22076						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	
Intercept	44.03102623	7.588395135	5.802416116	2.69781E-05	27.94434717	60.11770529	27.94434717	
X Variable 1	0.101352654	0.015587997	6.501967666	7.29665E-06	0.068307576	0.134397733	0.068307576	*Upper 95.0%*
								60.11770529
								0.134397733

Table 2: Simple Linear Regression of h-index with the total number of documents

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.875351198							
R Square	0.76623972							
Adjusted R Square	0.751629702							
Standard Error	17.95914103							
Observations	18							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	16915.50805	16915.50805	52.44618762	1.96779E-06			
Residual	16	5160.491947	322.5307467					
Total	17	22076						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	42.06961391	7.09816377	5.926830555	2.12768E-05	27.02217892	57.1170489	27.02217892	57.1170489
X Variable 1	0.001083856	0.000149663	7.241974014	1.96779E-06	0.000766584	0.001401127	0.000766584	0.001401127

Table 3: Simple Linear Regression of h-index with the total number of citations

We have also used simple linear regression to find the dependence of the number of citations and the number of publications on the value of the h-index.

7. Discussion and Conclusion

A look into the graphical representation of the dependence of the h-index with the number of documents published by the Nobel Laureates indicates that a higher number of publications has been associated with a high value of h-index and vice-versa. Similarly, the graph portraying the relationship of the h-index and the number of citations also show the same results. This leads to the conclusion that both these factors play a significant role in the determination of the h-index. The number of publications and the number of citations, both play a significant role in determining the value of the h-index. The value of $p<0.001$ in the simple linear regression shows that the number of publications and the number of citations are vital parameters while calculating the h-index. The results of our study confirm with the results obtained by Mahmoudi et al[18]. However, the coefficient has been calculated at 87.53% for the total number of citations, the value stands at 85.17% for the total number of documents. Though both the nondependent variables have a strong correlation with the h-index, the number of citations is a slightly better fit than the number of documents

published by the Nobel Laureates. The coefficient of correlation also agrees with our findings.

Khurshid et al (2018) have observed that in the h-index calculated using Google Scholar, the number of citations had a significant correlation ($r= 0.77$, $df = 113$, $p<0.0001$) with the h-index. Similar results were also observed in calculating the h-index using Scopus. Along similar lines, Hamidreza et al (2021) analyzed the h-indexes using various databases like the Web of Sciences, Scopus, and Google Scholar and found that the number of citations is strongly correlated with the h-index. In his proposal aimed at assessing the quality of a researcher based on the level of productivity, J E Hirsch (2005) proposed the h-index which takes into consideration the number of citations received by the scientific production of individual researchers. At the last but not least it can be concluded that the findings of this study indicate the fact that both the number of citations and the number of scientific production have an effect on the value of the h-index. Though extant studies have observed the correlation of the number of citations on the h-index, the present study is perhaps the first attempt at assessing the correlation between the number of scientific productions and the value of the h-index.

8. Limitations and Further Studies

This study has assessed the correlation of the number of citations and the number of products with the h-index with data extracted from the Scopus database. The use of articles in English without considering articles in other languages is one of the limitations of this study. Further, the use of the Scopus database with disregard to other databases is also a limitation. Further studies should explore other parameters like the number of times any scientific production is read, the number of times any researcher is invited to conferences, etc to predict the h-index. Future studies should also focus on calculating the relation of the h-index with other measures like m-quotient, e-index, g-index, i-10 index, etc.

References

1. Hirsch J. (2005). An index to quantify an individual's scientific research output. *Proceedings Of The National Academy Of Sciences*, *102*(46), 16569-16572. https://doi.org/10.1073/pnas.0507655102

2. Hirsch J. (2007). Does the h index have predictive power?. *Proceedings Of The National Academy Of Sciences*, *104*(49), 19193-19198. https://doi.org/10.1073/pnas.0707962104

3. Radicchi F., Fortunato S. & Castellano C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings Of The National Academy Of Sciences*, *105*(45), 17268-17272. https://doi.org/10.1073/pnas.0806977105

4. Henzinger M., Suñol J. & Weber I. (2009). The stability of the h-index. *Scientometrics*, *84*(2), 465-479. https://doi.org/10.1007/s11192-009-0098-7

5. Panaretos J. & Malesios C. (2009). Assessing scientific research performance and impact with single indices. *Scientometrics*, *81*(3), 635-670. https://doi.org/10.1007/s11192-008-2174-9

6. Abbott A., Cyranoski D., Jones, N., Maher B., Schiermeier Q. & Noorden R V. (2010). Metrics: Do metrics matter?. *Nature*, *465*(7300), 860-862. https://doi.org/10.1038/465860a

7. Acuna D., Allesina S. & Kording K. (2012). Predicting scientific success. *Nature*, *489*(7415), 201-202. https://doi.org/10.1038/489201a

8. Hamidreza K., Javad A., Ramin S. & Leili Z. (2013) H-indices of Academic Pediatricians of Mashhad University of Medical Sciences. *Acta Informatica Medica*, *21*(4), 234. https://doi.org/10.5455/aim.2013.21.234-236

9. McNutt M. (2014). The measure of research merit. *Science*, *346*(6214) 1155-1155. https://doi.org/10.1126/science.aaa3796

10. Yong A. (2014). Critique of Hirsch's Citation Index: A Combinatorial Fermi Problem. *Notices Of The AMS*, *61*(9)., 1040-1050.

11. Wilke C. O. (2014). Relationship between h index and total citations count. *Academia*, https://clauswilke.com/blog/2014/12/08/relationship-between-h-index-and-total-citations-count/.

12. Hicks D., Wouters P., Waltman L., de Rijcke S. & Rafols I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429-431. https://doi.org/10.1038/520429a

13. Ioannidis J., Klavans R. & Boyack K. (2016). Multiple Citation Indicators and Their Composite across Scientific Disciplines. *PLOS Biology*, *14*(7), 1-17. https://doi.org/10.1371/journal.pbio.1002501

14. Sinatra R., Wang D., Deville P., Song C. & Barabási A. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312), 596. https://doi.org/10.1126/science.aaf5239

15. Khurshid K., Shah S., Ahmadi M., Jalal S., Carlos R., Nicolaou S. & Khosa F. (2018). Gender Differences in the Publication Rate Among Breast Imaging Radiologists in the United States and Canada. *American Journal Of Roentgenology*, *210*(1), 2-7. https://doi.org/10.2214/ajr.17.18303

16. Ayaz S. & Masood N. (2020). Comparison of researchers' impact indices. *PLOS ONE*, *15*(5), 1-15. https://doi.org/10.1371/journal.pone.0233765

17. Koltun V. & Hafner D. (2021). The h-index is no longer an effective correlate of scientific reputation. *PLOS ONE*, *16*(6), 1-16. https://doi.org/10.1371/journal.pone.0253397

18. Mahmoudi M., Rahmati M., Mansor Z., Mosavi A. & Band S. (*2021*). A Statistical Approach to Model the H-Index Based on the Total Number of Citations and the Duration from the Publishing of the First Article. *Complexity*, 1-8. https://doi.org/10.1155/2021/6351836