PRZEGLĄD BIBLIOTECZNY 2024 z. 1
PL ISSN 0033-202X

VESLAVA OSIŃSKA
Institute of Information and Communication Research
Nicholaus Copernicus University in Toruń, Poland.
e-mail: wieo@umk.pl
ORCID 0000-0002-1306-7832

BERNARDETA IWAŃSKA-CIEŚLIK
Institute of Social Communication and Media
Kazimierz Wielki University, Bydgoszcz, Poland
e-mail: biwanska@ukw.edu.pl
ORCID 0000-0003-1841-6162

JAKUB WOJTASIK
Doctoral School of Social Sciences
Nicolaus Copernicus University in Toruń, Poland
e-mail: jwojtasik@doktorant.umk.pl
ORCID 0000-0001-6157-5658

BRETT BUTTLIERE
Center for European and Regional Studies (EUROREG)
University of Warsaw
e-mail: brettbuttliere@gmail.com
ORCID 0000-0001-5025-0460

JOANNA KARŁOWSKA-PIK
Faculty of Mathematics and Computer Science
Nicholaus Copernicus University in Toruń
e-mail: joanka@mat.umk.pl
ORCID 0000-0001-9157-7355

ADAM KOLA
Faculty of Humanities
University Centre of Excellence IMSErt – Interacting Minds, Societies, Environments Institute for Advanced Study, Nicolaus Copernicus University in Toruń, Poland University of Amsterdam, Amsterdam, the Netherlands
e-mail: adamkola@umk.pl
ORCID 0000-0002-0584-6342

SCIENTISTS' CONTRIBUTION TO THE IDUB RANKINGS. POLISH RESEARCHERS ON THE GOOGLE SCHOLAR PLATFORM

1. Veslava Osińska
Veslava Osińska –is associate professor at the Institute of Information and Communication Research at the Nicolaus Copernicus University and a principal investigator of the Polish team in the international Chist-era project – Bitscope (bitscope.umk.pl). Her interests are multi-scale data visualization methods, in particular science visualization. She is a lector of subjects related to data processing, analyses and visualization.. Veslava Osińska is a member of several societies, both national and international: Polish Information Technology Society, International Society of Knowledge Organization and the Association of Polish Scientists in Lithuania.

2. Bernardeta Iwańska-Cieślik
Bernardeta Iwańska-Cieślik – Ph.D., assistant professor at the Department of Journalism and Media Research at the Institute of Social Communication and Media at the Kazimierz Wielki University in Bydgoszcz. Her research interests revolve around the history of books and the press in Włocławek, and she also deals with bibliometric issues based on the publishing activity of academic librarians in the field of old books and the press. Author of the book Biblioteka kapituły katedralnej we Włocławku (2013), editor of collective works and several dozen scientific articles, including: Informacja o nowych publikacjach polskich bibliologów i informatologów w przestrzeni sieciowej („Toruńskie Studia Bibliologiczne" 2016).

3. Jakub Wojtasik
Senior data analyst at the Center for Statistical Analysis at the Nicolaus Copernicus University in Toruń. He is a PhD student at the Doctoral School of Social Sciences in Nicolaus Copernicus University in Toruń. Fellow of the Polish National Science Center and the Polish National Agency for Academic Exchange.

His research interests include issues of mathematical modeling, applications of data mining methods and machine learning in economic models and forecasting, as well as optimization theory.

4. Brett Buttliere

Brett Buttliere – Ph.D., works at the Center for European and Regional Studies (EUROREG) at the University of Warsaw. Author of several papers across areas such as psychology, bibliometrics, psychology of science, and communication, his interests mainly came from an understanding that science is done by humans, and that any problems and potential solutions must consider this humanness. He has worked at universities in the United States, the Netherlands, Germany, and Poland, and contributed to conferences across the world. He has variously surveyed scientists about open science (2014), analyzed conflict in scholarly tweets and article keywords (2017), synthesized 'alternative' metrics of impact (2017), studied science on Wikipedia (2021), outlined mechanisms enabling shareable analysis scripts (2021), and developed the meta.data() R package (2023). He is actively working on encouraging scientists to engage with and contribute to Wikimedia, encouraging academic societies to host conferences in developing nations, developing more sustainable and creative research environments, and developing a more general and applicable model of minds.

5. Joanna Karłowska-Pik

Joanna Karłowska-Pik, assistant professor at the Faculty of Mathematics and Computer Science, Nicolaus Copernicus University in Toruń, and Director for the Centre for Statistical Analysis, NCU. Holds a PhD in mathematics. Trainer of IBM SPSS Statistics Software. Research interests and expertise include stochastic processes, statistics and data science – mainly applications of machine learning in medicine and natural sciences.

6. Adam Kola

Adam F. Kola is a director of the Center of Excellence IMSErt: Interacting Minds, Societies, Environments and associate professor at Nicolaus Copernicus University, Toruń, Poland. In 2021-2022 he was a visiting researcher at the Institute for Advanced Study, University of Amsterdam; in 2016-2019 he was a visiting scholar at the University of Chicago. His research has been focused on Eastern and Central European intellectual and literary history, memory studies of the 19th and 20th centuries, and global knowledge transfer. His most recent books are: 'Studying the Memory of Communism. Genealogies, Social Practices and Communication' (eds. with R. Halili, G. Franzinetti, 2021, in English) and 'Socialist Postcolonialism. Memory Reconsolidation' (2018, in Polish). He is the author of about 100 papers in Polish, Czech, Russian, German, and English, and he translates Czech and English into Polish.

KEYWORDS: Google Scholar. Scientific literature. IDUB

ABSTRACT: **Thesis/Objective** – Google Scholar is a tool that is widely used not only to search the scientific literature, but also to obtain information on researchers' scientometric measures. In this article, we will verify whether, based on GS data, users with the highest measures will be identified as associated with the best universities in Poland, called IDUBs. **Methodology –** Stepwise logistic regression models with cross-validation were used to find variables significantly influencing the correct automatic classification. **Findings and conclusions** – The best models in terms of predictive quality were obtained using the h-index, the type of university, the annual number of publications and the year of the first publication as predictors. Student's t-tests showed statistically significant differences in the mean values of the h-index, the i10 index and the number of publications (p<0.001, p<0.001 and p=0.013, respectively) between researchers from the best 10 universities in Poland (associated as IDUBs) and scientists from other academies. The scholars characterized by high scientometric measures were affiliated to IDUB schools – this relationship is observed within the scope of universities, not technical or medical schools. Due to the free and open nature of the GS, the data obtained from it are heterogeneous and often incomplete, making automatic processing and analysis difficult. These complications are particularly evident when aggregated rather than individual data being analysed. Despite these limitations, the results obtained make it possible to cope with the rapid growth of scientometric data and may lead to the creation of new measures for assessing the scientific output of scientists.

1. INTRODUCTION

In many cases, Google Scholar (GS) is the first source of bibliographic data on a given topic, because access to its resources does not require special search skills. GS also offers functionalities that allow to build one's own information space, including the ability to follow a topic. Researchers' profiles provide a list of individual works and information about the popularity of those works (e.g., the number of times they have been cited), along with personal bibliometric indicators. Given the prominence of this information, it can be considered significant in a particular research scope. GS, itself, suggests top studies and scholars for both individuals and research fields more generally. GS, the free academic search engine, is only one option in the search for scientific literature, and its metrics yield only indicative data (Harzing, 2017; Gusenbauer & Haddaway, 2020). However, the data is becoming an increasingly significant element in the evaluation of the achievements of individual scientists.

The majority of records in the GS database, at least relating to researcher profiles, is driven by the user or generated automatically, where the ease of use and functionality of the platform may encourage the researcher to keep a profile updated. This is important, because if the individual does not create a profile, their citations are not collated, whereas in the imposed or official metrics, these are done by the companies to the best of their abilities (Google Scholar Profiles, 2021). Taking this into account, we can say that GS data is highly dependent not only on scholar activity in terms of publishing but also on a scholar's willingness to create and curate own GS accounts. Curation refers to tracking and monitoring updates, checking the correctness of bibliographies, searching for researchers with similar interests, and subscribing to recently indexed articles. In this sense, the GS database is mainly oriented towards previewing a researcher's achievements, not institutions. This leads to the conclusion that, in the context of institutions, statistics combined from aggregated GS data for each institution may differ.

It was already observed that GS indicators such as citations and the h-index remain the most-used metrics of a scholar's impact because of their ease of access (López-Cózar et al., 2012; Google Scholar Metrics, 2021). However, their applicability is questioned, especially at a national level, and essential differences exist from values obtained using the Web of Science or Scopus (Bar-Ilan, 2008; Moed et al., 2016; Martin et al., 2021). Nevertheless, countries, such as the UK and Australia, use GS, and citation data it collects serve as an additional metric for performance evaluation and building rankings for 130 universities (Mingers et al., 2017). Scientists' data available in GS have become one of the elements of the assessment of scientific units and their employees, who apply for promotion or further employment (Harzing & Alakangas, 2016; Bornmann et al, 2016; Prins et al., 2016; Jensenius et al. 2018). In Poland, GS citation counts have been taken into account when considering national grant applications.

THE SELECTION OF IDUBS

In 2019, the Polish Ministry of Science and Higher Education conducted the first "Initiative of Excellence – Research University" (IDUB) program in the context of the new Constitution for Science (IDUB, 2019). It aims to select and support universities that will strive to achieve the status of a research university and be able to effectively compete with the best academic centres in the world. The universities participating in the competition were evaluated according to established rules. The methodology is based on the assessments of a series of predefined parameters and weighted accordingly. This is, first of all, the scientific efficiency of the institution, and its internationalisation, innovativeness and prestige. Graduates' success in the labour market, as well as the condition of education, were also taken into account (Prawo, 2021). Thus, this ranking selected IDUB institutions in Poland for the period 2020–2026, where both quantitative and qualitative indicators were considered (Komunikat 2018; IDUB 2019). For reference, the years for the calculation of scientometric indicators from the period

2013–2017 were taken into consideration, but the selection of data sources (Web of Science or Scopus) was dependent on a particular university.

Since 2020, ten Polish academic institutions have been qualified to IDUB. Among them, there are five universities (according to the traditional Latin meaning, *"universitas magistrorum et scholarium"*), four universities of technology and one medical academy. Table 1 presents these institutions, their full names, short forms which will be used in the next sections and their rankings from 2021: Scopus and Leiden rankings. Scopus measures relate to various combinations of scientometric indicators of particular researchers indexed in the database, whereas the Leiden Ranking ranks universities worldwide between 2016 and 2019 by the number of academic publications according to the volume and citation impact of the publications at those institutions (Methodology, 2014; Waltman et al., 2012; Waltman & van Eck, 2013). It is important to note that Leiden rankings are based on data derived from the Web of Science. Thus, the composition of Table 1 is intended to be an initial characterisation of the selected top ten universities using the two main global databases.

Table 1. The ten best Universities in Poland according to the 2021 ranking of Polish universities

No	University	Scopus rank[1]	Leiden rank[2]
1	Warsaw University (UW)	**2**	**2**
2	Gdańsk University of Technology (PG)	**10**	**12**
3	AGH University of Science and Technology (AGH)	**3**	**3**
4	Warsaw University of Technology (PW)	**4**	**5**
5	Adam Mickiewicz University (UAM)	**7**	**4**
6	Jagiellonian University (UJ)	**1**	**1**
7	Gdańsk Medical University (GUM)	**21**	**23**
8	The Silesian University of Technology (PS)	**6**	**11**
9	Nicolaus Copernicus University (UMK)	**8**	**8**
10	Wrocław University (UWr)	**9**	**13**

[1] According to scholarly output – citation count dated 5 September 2021.
[2] at 2021.

The selection was performed based on both the general achievements of scholars and the visibility of these institutions' research in global science. However, this evaluation was developed only on the basis of scientometric databases such as the Web of Science or Scopus, which were more focused on an institution's achievements. In the context of the aforementioned popularity and accessibility of GS, it seems reasonable to check whether the data it contains allows similar ratings to be given to academic institutions, as well as the relationship between official institutional statistics and individual scholar data on GS. What is more, by examining scholar visibility on GS under their institutional domains, we seek to shed light on the reliability of GS metrics in representing academic communities.

By analysing the above data and that obtained from the GS platform, the authors formulate several hypotheses:

H1: GS data in the context of individual scholar achievements corresponds with the official ranking of IDUBs in relation to other national scientific institutions.

Analogous to IDUB items, the whole database was grouped into three types of universities. The final comparison of items between and within the groups seems to be more reasonable and easier to draw conclusions from. Therefore, a second hypothesis is proposed as follows:

H2: There are significant differences among the representation of active scholars on GS in relation to the different types of institutions, universities, technical and medical.

It should be noted that GS procures an overview of citations in each indexed publication. Such important bibliometric information is missing from Polish bibliographic databases, and is expected to continue to be so. GS remains a general and easily accessible source of citation scores as a basic and quick parameter of an article's popularity and related authors' recognition within the community. Indeed, the GS database contains more bibliometric indicators, and they will all be used in the current study. However, citation rankings can also be analysed through the modification of the collected database; for example, by reducing selected records. The next hypothesis concerning this approach is thus:

H3: Removing records with extreme values from the GS dataset (scholars with either very low or very high citation counts) should neither change discovered dependencies nor effect conclusions.

From the perspective of the sociology of science, it is interesting to identify the groups with no or low impact. Therefore, an auxiliary hypothesis will be:

H3A: The mentioned records of low citation impact being removed from the GS dataset mainly refer to students or randomly created accounts.

2. GOOGLE SCHOLAR PLATFORM

GS gives us access to the greatest scientific resources in the world (Gudenbauer & Haddaway, 2020). In January 2018, its resources were estimated to contain references to 389 million documents (ibid). The available scientific materials include journals and books, conference papers, theses and dissertations, preprints, abstracts, technical reports, and other scholarly literature, including court opinions and patents, as well as grey literature and full texts.

From the very beginning, the authors undertook a comparison of the number of available publications and their citations between such giants as the Web of Science and Scopus with GS, and first was Peter Jacsó (2008, 2012). Kiduk Yang and Lokman I. Meho (2007) also analysed the number of citations and access to more publications. A search for alternatives to the Web of Science was also undertaken to find more articles in the field of social sciences. GS was also indicated as an alternative, although, in the end, it was not treated seriously as competition for WoS (Norris & Oppernheim, 2007).

Michael Gusenbauer undertook a complete comparison of scientific search engines, initially comparing the sizes of 12 academic search engines, and proved that GS has no competition in this field (2019a). In another study, he compared 23 search engines (2019b), and in collaboration with Neal R. Haddaway, they put together 28 sources in which scientific publications were catalogued. The list of sources, among others, included GS, PubMed, WoS, EbscoHost, Microsoft Academic, Scopus and Springer Link. Most of the proven brands were considered to be the main sources of publications, while GS was found to be effective only for supplementing bibliographic searches. The skills a researcher should have to efficiently navigate in very different systems were also emphasized (Gusenbauer & Haddaway, 2020). The GS number of citations and the h-index became a pretext to think about the index itself and search for influential scientists from selected fields or the presence of representatives of individual universities. Erroneous h-index indications were reported by Jaime A. Teixeira da Silva (2018).

There is a clear interest in GS from medical scientists, who initially compared PubMed and GS resources (Shultz, 2007). Then, the resources of PubMed/MEDLINE, ScienceDirect, Scopus, and GS were analysed in terms of publications on laser medicine, and the discussed search engine was the most effective (Tober, 2011). Recently, it has been hypothesized that GS is one of the main resources to search for the latest medical publications (Anders & Evans, 2010; Bramer et al., 2013). One of the hot topics in scientometrics concerns analyses of resources on academic platforms (so-called social media for scientists), such as Academia.eu and ResearchGate (Thelwall &Kousha, 2017). The authors answered the question of which platform found more early citations. Upon analysis, it was found that RG is not yet able to compete with the indexing capabilities of GS.

Metadata from GS describing individual scientific publications are burdened with numerous errors, which has been noted in a recent article of a researcher from France – Romy Sauvayre (2022). However, it still gives a real possibility of recognizing research, especially representatives of the humanities and social sciences, on the international arena. The lack of publications in the field of humanities and social sciences in WoS and Scopus was noticed very quickly. Anne-Wil Harzing, the founder of Publish or Perish, published her observations in a blog, *Google Scholar is a serious alternative to Web of Science* (2017). Her conclusion was that: "Google Scholar and Publish or Perish have democratised citation analysis".

One of the greatest achievements of GS creators was the introduction of the possibility to build your brand by setting up a private profile in Scholar Citations Profiles. The benefits of creating a profile include the ability to group publications in one place under an appropriate name and increase the visibility of scientific achievements. GS has also become important for universities, and on practically every university website you can find tips on how to increase your visibility in the Web space, including creating a GS profile (Bogajczyk, 2019). In 2013, in Poland, Emanuel Kulczycki (2013) prepared a guide for scientists on how to create a GS profile and add new publications to it.

Polish researchers began examining the activity of scientists in self-representing in alternative channels to official national bibliographies, institutional bases of publications and sometimes very young repositories. The group which was the most frequently analysed were representators of communication science and media, especially scientists of information science (Świgon et al. 2022). Among them, the most popular place to mark your presence in the world of science was Google Scholar Profiles. Further places were taken by Academia.edu and ResearchGate. Five years earlier, Hołowiecki (2017) noticed the opposite proportions, indicating that Poles were more interested in the Academia.edu portal. In turn, in 2015 Pulikowski tested whether Polish articles are visible in Google and Google Scholar, Bing and Base. Publications from Polish repositors and digital libraries are well recognized by Google search engines (2015).

In the presented study, it is important not only to obtain information on the number of citations, but also increase the recognition of Polish research and its visibility in the network space. In the previous research studies on the rapid dissemination of scientific texts, comparisons included, apart from WoS, Scopus or PubMED, primarily GS and social media for scientists. Such a comparison was made many times showing that GS is one of the strongest medium to present latest works (Dorsch, 2017). Recently, one of the most important elements in scientific communication is the promotion of articles and other works (D'Alessandro, 2020). This article, for the first time, will present all Polish scientists who decided to create a profile in Google Scholar – an academic search engine and not only contrast this with the national list of schools of excellence but also show their contribution at present in this noble list.

3. DATA AND METHODS

3.1. BASIC CHARACTERISTICS OF THE DATASET

The data for this study were collected using the Python scraping library BeautifulSoup from the GS platform over the first quartile of 2021. The final dataset consisted of 28,375 records. Using the collected dataset as a basis, we next used the R environment to build the final database. The ScholaR package was used to scrape the profile of an individual scientist based on previously collected Google Scholar IDs (Yu et al., 2021). The procedure makes available the individual's name, affiliation, the total number of citations, the h-index, the i10 index, the field of work, and any link to their homepage. In the next step, using the same R package, additional data such as the overall number of articles and the earliest publication year were collected and processed. After removing duplicates and performing data cleaning, the number of records analysed was $N = 20{,}751$.

Ten schools of excellence (IDUB) represent three types of academic institutions in Poland: universities, the largest teaching profile, polytechnics (technical universities) and medical universities. Data was gathered from the GS institutional accounts of each of the three types of schools. Only those schools that were able to be found by scraping the GS space were considered, resulting in a collection consisting of 18 universities, 17 polytechnics and 9 medical universities. Thus, 44 institutions were chosen and compared in terms of GS individual profiles. For clarity of future results, short forms of institutions' names were created. For universities and universities of technology short forms were prefixed with U and P, respectively (with the exception of the AGH University of Technology and Science, where the commonly used AGH short form was used). The rest of the short forms were built from the name of institution. By analogy short forms were created for medical universities, but instead of the prefix, -UM, suffixes were added. The table with full names and short forms is available in the appendix.

Another data source was required to receive information about the number of hired researchers in the aforementioned institutions. There are several databases relevant to Polish science. Nauka Polska ("Polish Science") is the oldest database of the National Information Processing Institute, having been developed since 1990. The dedicated platform (http://nauka-polska.pl) stores and maintains resources relating to scientific and R&D publications, doctoral dissertations, habilitation theses and expert reports. However, another academic database, Radon, has been gaining importance in recent years. This is a knowledge-based platform (http://radon.gov.pl) providing the most reliable data on Polish science, and built-in tools for reporting and visualising. It relies on the modern system of scientific information management, which imports data from multi-distributed sources, among others conducted by the National Information Processing Institute (OPI-PIB). Due to more current data, the Radon database was selected to reach the number of academic teachers employed at a particular university.

By composing the two main quantities collected from GS and the Radon database, we can estimate the ratio of scholars from every university knowing about and engaging in the GS network. Table 2 presents the data sorted in descending order according to the percentage of researchers with an account on GS from each university. Therein, a large discrepancy in the number of scholars' accounts is seen, ranging from 17 (Jan Kochanowski University in Kielce) to 2187 (Warsaw University). However, these disparities may be overestimated. It is common practice to combine work in multiple research centres. Thus, smaller (and implicitly less prestigious) universities may be left out of the introduced affiliations.

Table 2. The quantities of researchers and GS accounts registered to the universities. IDUB universities are bolded. For full names of universities, see Appendix

Universities					Technical Universities					Medical Universities				
#	School	N of GS account	N of staff	Ratio	#	School	N of GS account	N of staff	Ratio	#	School	N of GS account	N of staff	Ratio
1	US	1188	1921	0.62	19	**PW**	**1602**	**2502**	**0.64**	36	WUM	220	1845	0.12
2	UG	1079	1810	0.60	20	**AGH**	**1178**	**2191**	**0.54**	37	**GUM**	**142**	**1148**	**0.12**
3	**UW**	**2187**	**3959**	**0.55**	21	PP	714	1335	0.53	38	LoUM	137	1598	0.09
4	**UMK**	**1137**	**2362**	**0.48**	22	PWr	1082	2141	0.51	39	PUM	121	1574	0.08
5	**UWr**	**810**	**1955**	**0.41**	23	PB	301	631	0.48	40	SUM	115	1421	0.08
6	**UJ**	**1588**	**4467**	**0.36**	24	**PS**	**767**	**1638**	**0.47**	41	WrUM	90	1345	0.07
7	UWM	598	1778	0.34	25	PCz	298	666	0.45	42	BUM	62	887	0.07
8	USz	346	1017	0.34	26	PSz	342	830	0.41	43	SzUM	48	696	0.07
9	UL	664	2248	0.30	27	PL	208	585	0.36	44	LUM	68	1394	0.05
10	UB	241	793	0.30	28	PLO	430	1212	0.35					
11	UKW	174	635	0.27	29	PK	388	1099	0.35					
12	UKSW	192	793	0.24	30	Pkie	144	428	0.34					
13	UO	206	890	0.23	31	PRz	280	930	0.30					
14	**UAM**	**603**	**2939**	**0.21**	32	**PG**	**405**	**1430**	**0.28**					
15	UMCS	314	1582	0.20	33	PO	112	450	0.25					
16	UZ	171	1046	0.16	34	PBB	82	347	0.24					
17	URz	101	1328	0.08	35	PR	72	354	0.20					
18	UKie	17	961	0.02										

A similarly large discrepancy was registered in the percentile coverage of GS in the researchers' community comparing the universities with each other. As many as 64% of scholars employed at Warsaw University of Technology have accounts on GS. Similarly, almost half of the scholars of UMK, UW, the University of Silesia (US) and Poznań University of Technology (PP) are visible in the GS space. On the contrary, there are also universities, mainly medical, that have 6 percent or less of GS users. It can be quickly observed from Table 2 that IDUB institutions are also represented on GS space unequally, with ratios between 0.12 and 0.64.

3.2. DATA COLLECTION AND PROCESSING

The GS table consists of the following fields:

- the name of researcher,
- a link to a picture,
- a link to a profile,
- an email address,
- a citation count,
- a description (depending on the schools' template),
- keywords (depending on the schools' template).

Some of these fields remain unfilled, but the researcher's name and link to his or her profile are obligatory on the GS platform. The link to a profile is constructed in such a way that it enables the extraction of the researcher's ID from the GS database.

The authors collected data in two phases: (1) filtering scholars' IDs and (2) scraping scholars' data from their profiles. During this process data many problems have been dealt with. Early searching and filtering data by a university domain revealed several unexpected difficulties due to changing university domains in Poland in 2010. Thus, the first step in searching was conducted separately for current and previous domains and the different language versions of university names. Another issue arose in connection with the number of co-authors, because the Scholar library does not deliver data about all co-authors if a researcher does not accept their list in the GS environment. It was misleading to see prominent scientists with high citation rates having not a single co-author, whose names should appear in a separate, righthand panel. It is also unclear how and whether the individual threshold for the co-authors number was set up by the R package. Thus, the number of collaborators will be treated as an uncertain variable. What is more, the collected years for the oldest article were not reliable for the whole dataset, as two- and three-digit numbers occur, as well as the range 1700-2021. The reason for this is that some individuals do not care about providing correct metadata and, for example, enter "80" instead "1980". Therefore, the minimal threshold for credible years for our analysis was established on 1960. There were also plenty of individual problems to solve, such as encoding, using external mail addresses or ones with a European domain, adding a scientific degree to the name field or multiple combinations of the name for each scientific institution.

Using the collected dataset (has been uploaded it into open ICM repository at address: https://doi.org/10.18150/PGS2H8) as a basis, we next used the R environment to scrape data. The data was scraped from GS using the ScholaR package in R. This package, which may be downloaded from https://cran.r-project.org/web/packages/scholar/scholar.pdf, makes a wide variety of functions available that plug into Google API, including functions to obtain the citation history of a paper or person, their profile, to create a plot of co-authors and many more functions (Yu et al., 2021). To clean the university variable, the data was cleaned using the greps and gsubs functions, which essentially identify a character string and then replace it with another. The first steps involved removing any indication of (e.g.) "dr", "prof" and "adjunct", which are often splitters. Next, the data was examined to identify common ways of expressing the university name.

Regarding the representativeness of GS data, as the data was scraped from a website, we also took into consideration inactive users (e.g., sham or unused accounts). For example, retired researchers and staff or graduated students might keep their profiles, rendering data out of date. But exact data about how many accounts were unused was difficult to obtain, if not impossible. However, the authors can claim that by estimating the set of UMK accounts, the calculated ratio should be reduced by roughly 15–25 percent.

The detected unreliability of data is implied by internet origin, which influences the deviation from a known distribution such as normal, or log normal (Thelwall, 2013; Buttliere & Buder, 2017). Based on the assumption that errors occur evenly or close to evenly across institutions, this predisposes the conducting of further comparative studies between academic schools.

As one of the results, the 20 most-cited scholars are presented in the Annex Table. Examining this table in detail demonstrates the potential downsides of GS data, as it is significantly less clean than other data. For instance, "Bartek Lipinski" is rated the fourth-most-cited scholar in Poland, but his email address points to student status at UMK and he is not on any of the top papers he has attributed to him. This is problematic, as GS does not sometimes properly index documents due to the incorrect identification of names/surnames of authors. Another problem is the presence of misleading or no institutional names of schools, such as in the two cases (shaded background) in the Annex. We tracked the entire data table manually and specified 11 records that can be qualified as uncertain in relation to the name of the scholar, their affiliation or email address. It should be mentioned that institutional email addresses are a reliable qualifier of the analysed values. If the email address belongs to an educational institution, the table returns the status "verified email at …"; in other cases, "no verified email".

Still, the Annex is interesting in other ways: one might expect that the top researchers would be the main drivers of the effect between IDUB universities and the rest of schools. However, only eight of the top 20 researchers are affiliated with IDUB universities.

3.3 STATISTICAL ANALYSES

Data from researchers affiliated to IDUB and non-IDUB scientific institutions were compared. Basic descriptive statistics were made to summarize the characteristics of selected subsets. Variables that were studied should, of course, directly relate to scholars' achievements. Thus, we focused on the citation count, standardized citation, the h-index and i10-index (own GS indicator with a similar principle as the h-index, measuring the number of publications with at least 10 citations) by GS. A comparison of the means of selected variables concerning the two groups were performed with Student *t*-tests. Distribution was analysed using both histograms and Q-Q plots. Additionally, the original dataset was grouped by university type, and the basic statistics were accomplished. If the aggregation level provides a macro view of the dataset, the particular records-based pattern can reveal essential details about both individuals and new groups of data.

In the current research, we referenced Bornmann's studies (2016), which described how adding new variables to explain the differentiation of the schools can improve each model. They used a logistic regression model for institutional bibliometric evaluation. Their purpose was focused on the question of whether it is possible to predict excellent schools based on citations or citation-based indicators. Therefore, for institutional comparisons, one needs to construct an appropriate model that relies on various sets of working variables determining the scientific impact. Thus, stepwise logistic regression models with cross-validation were used to find variables significantly influencing the correct classification.

For logistic regression modelling, model performance estimation we used the PS Imago PRO 7 (based on the IBM SPSS Statistics 27 analytical engine). Statistical tests, descriptive analysis, visualizations and insights were performed in Python (version 3.8.10) with additional libraries: pandas (1.4.3), matplotlib (3.1.3), scipy (1.10.1), and seaborn (0.11.2).

4. ANALYSIS OF RESULTS

4.1. AGGREGATION LEVEL

According to the initial assumption, GS data allows the viewing, as well as analysing, of individual accounts. However, statistical properties of aggregated groups are required to select the characteristics that distinguish them most. The basic statistics of GS accounts (as obtained in the initial step of the analysis) regarding the type of analysed institution – medical, technical and general university – are shown in Table 3.

Table 3. Basic statistical indicators of Polish universities according to the type of institution

Type of institution	Number of institutions	Number of scholars	Mean institution size	Citations mean	Sum of citations	h-index mean	i10 index mean	Number of scholars with co-authors list	Mean number of co-authors
Medical	9	985	109.44	1,403.65	1,382,592	10.82	18.17	273	5.15
Technical	17	8,338	490.47	409.33	3,413,032	7.43	8.68	3,412	5.30
Universities	18	11,428	634.89	396.27	4,528,581	6.56	7.74	3,739	8.03

Despite basic descriptive statistics, such as mean values or sums, we computed average institution size by dividing the number of individual scholar accounts by the number of institutions of a particular type. With such a variable it is clearly seen that Polish universities are better represented in GS (634.89 accounts, the average for an institution) than technical (490.47) and, especially, medical ones (109.44), what is consistent with previously obtained ratios (compare Table 2) However, in spite of the disparity in citation totals to the detriment of medical schools, there is a noticeably higher average citation value for scientists in this group of universities. It is caused by the fact that they have many more publications with at least 10 citations (i10 index) than scientists from other types of universities. An inverse relationship is observed for co-authorship statistics. A common pattern in global science is that the number of co-authors on average is highest in medical science teams and lowest in the humanities and social sciences (Wang & Barabási, 2021). However, this is not evident from Table 3, where the highest average value is observed for the group of universities. This is undoubtedly influenced by the observed lower proportion of medical school researchers with GS accounts, which prevents the full identification of co-authors. Other reasons for the disproportion are non-complementariness (a small number of authors with co-author lists) or unreliability (counting the threshold in the R library) of GS data.

Another important factor which needs to be considered is the representativeness of the total population of Polish researchers by university type. A comparison of the frequencies in two databases, GS and Nauka Polska, is presented in Table 4. The quantitative proportions between these three categories are 1: 2.6: 5.8 in the case of available databases () and 1: 8.4 : 11.6 from the GS data (experiment). If we combine these numbers, one can note that medical schools are largely under-represented by scholars on GS.

Table 4. The number of registered employees reached from Nauka Polska and GS

Type of institution	Values from Nauka Polska	Values from GS
Medical	17,321	985
Technical	44,541	8,338
Universities	100,808	11,428

To better present the differences between citation measures, data was standardized by subtracting the mean value of each variable and dividing it by its standard deviation. Thus, transformed variables have a mean value equal to zero and a standard deviation equal to 1, whereas their original distribution remains. Such a modification is informative in terms of interpretation in separated types of institutions, which is shown in Table 5. Positive values are interpreted as greater than the mean value of all researchers, whereas negative – smaller.

Table 5. Standardized basic GS indicators

Type of institution	Number of institutions	Number of scholars	Standardized citations	Standardized h-index	Standardized i10 index mean	Number of scholars co-authors	Standardized mean number of co-authors
Medical	9	985	0.303	0.50	0.45	273	−0.24
technical	17	8,338	−0.013	0.04	0.00	3,412	−0.21
Universities	18	11,428	−0.017	−0.07	−0.04	3,739	0.21

For standardized data resemblance of mean values for all researchers and those affiliated to technical universities is observed. Additionally, we obtained a significant positive deviation for measures in medical universities. Standardized average numbers of collaborators for technical schools and universities are arranged equally around the mean, at −0.21 and 0.21.

Apart from the type of university, data can be grouped by IDUB assignment. For such division we compare, by analogy, descriptive statistics placed in Table 6. Even though the IDUB group only includes 10 institutions, almost half of the GS community collected in the dataset is affiliated to its constituent universities. This allows us to assume significant network activity in terms of the scholarship communication of IDUB scholars. What is more, using statistical tests, we can verify if the differences in scientometric variables observed in the data can be generalized to the whole population. Due to a large sample size, we can test variables with the t-Student's test without verifying normality assumptions (Elliott & Woodward, 2007).

Table 6. Descriptive statistics and results of the Student's t-test for the basic scientometric variables grouped by IBUD membership

Variables	IDUB UNIVERSITIES					NON-IDUB UNIVERSITIES					Student's t-test
	N	mean	standard error	median	IQR	N	mean	standard error	median	IQR	p-value
TOTAL CITES	10280	485,26	25,00	103	314	10471	414,07	35,85	95	270	0,104
H INDEX	10280	7,32	0,08	5	6,25	10471	6,90	0,07	5	6	<0,001
I10 INDEX	10280	9,12	0,21	3	9,25	10471	8,12	0,20	3	8	<0,001
NUMBER OF PUBLICATIONS	10276	49,78	0,66	31	50	10470	52,07	0,65	35	48	0,013

As the results show (Table 6) in the breakdown of data according to status, there are statistically significant differences in the means of the variables describing the h-index, the i10 index and the number of author publications. The lack of a statistically significant difference in the mean number of citations may be counterintuitive to the values observed in Table 5. However, test results depend on the value of the standard deviation of the variable, which is large due to the presence of outliers in the sample.

Executed statistical tests together with descriptive statistics proved that IDUB universities receive higher rates of, the h-index and i10-index ($p<0.001$ in both cases). We observe higher values of mean citation, but the statistical significance ($p=0.104$) of the test does not allow the generalisation of this observance. On the other hand, the number of publications is greater for non-IDUB institutions ($p=0.013$). Combining all this information can lead to the conclusion that researchers from IDUB institutions are leading in qualitive research and their publications are more influential. Thus, the main research question – do the best universities in Poland employ the best scholars – may have a positive answer.

4.2. INDIVIDUAL LEVEL

Scientometric measures distribution

The first step in analysing row data and matching statistical tests is usually to check the behaviour of data (i.e., how it is distributed). As befits statistical data derived from the internet and related to user behaviour, it is characterized by significant skewness: citation counts vary from zero to several hundred thousand (Thelwall & Wilson, 2014). It is commonly used to apply the logarithmisation of variables (citation count, h-index and i10-index) in the case of a strong skew profile (Buttliere & Buder, 2017).

By visual analysis of histograms of logarithmised variables (Fig. 1 A, B), it is possible to estimate how much the distribution deviates from normal. Another approach makes use of two numerical measures: skewness and kurtosis. (Orcan, 2020; Altman & Bland, 1995, 1996). To evaluate the distribution, the variables were scaled by increasing by 1, and then logarithmised.

The records with zero citations constitute only 3.4 percent of the entire dataset, which corresponds to the left bar on the histogram. From our observations, it can be assumed that these records are assigned to young researchers (PhD, students or assistants who are just starting their careers). However, based on autopsy, we can observe that the output of young

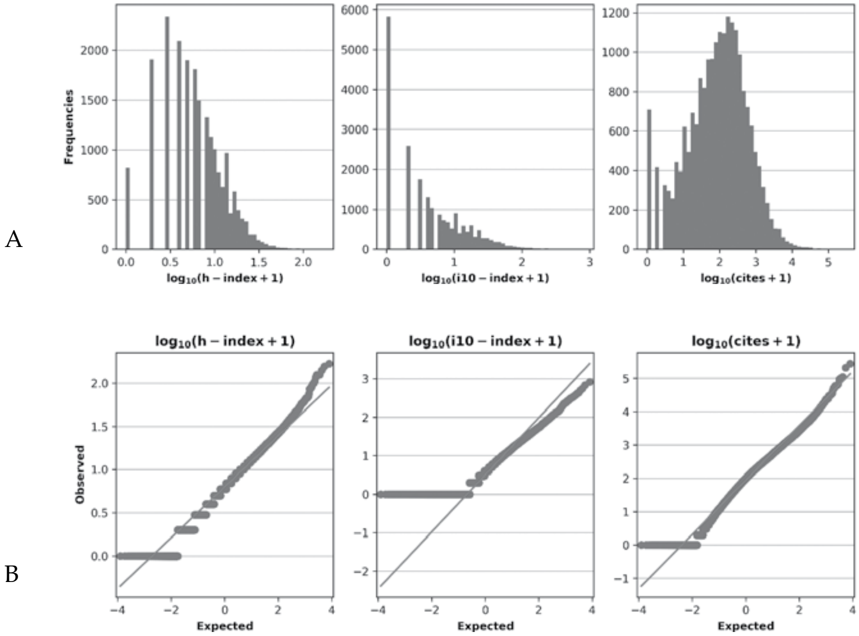


Fig. 1. The distribution of the logarithm of variables: h-, i10-indexes and citations (A) accordingly and Q-Q curves (B)

researchers depends strongly on the domain in which they work. For example, second year PhD students of physics can reach as many as 100 citations because of co-authorships with supervisors who may collaborate with prominent researchers. Conversely, some senior researchers in the humanities have only 20–40 citations on GS, which may be as a result of their printed works not being indexed. One of the hypotheses (H3) of the current research was to confirm that no differences occur by cutting the long tail, which probably constitutes negligible citations overall. It was difficult, if not impossible, to establish seniority status according to citations or the h-index count; therefore, more detailed studies should be undertaken to test the last hypothesis.

The h-index versus i10 index

GS implemented its own author-level metric, the i10-index, based on the same principle as the h-index (Teixeira da Silva, 2021) and defined as the number of publications with at least 10 citations. Thus, a minimum number of citations is predefined instead of the number resulting from quantitative relationships between publications and their citations. This measure of researcher productivity is more selective than the h-index, as evidenced in that the h-index can be higher than the i10-index, whereas

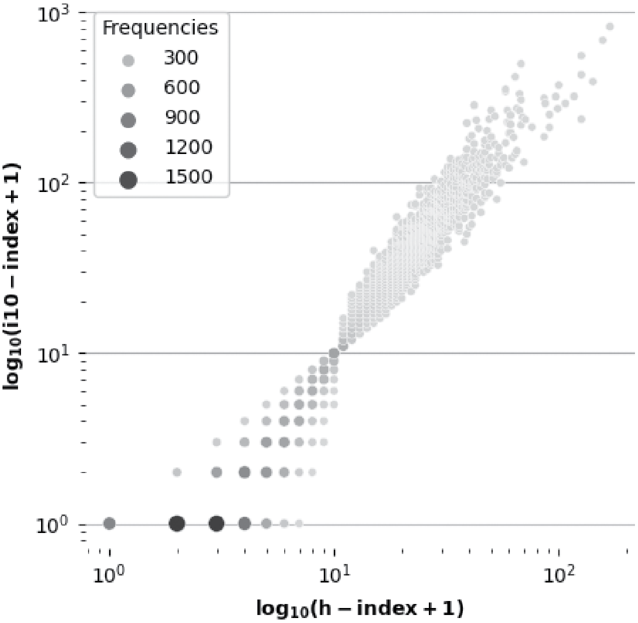


Fig. 2. h-index and i10 index dependence in log-log scale. The plot makes it possible to see that anything under 10 is very regulated. Note that the h-index can be higher than the i10-index, whereas the i10 cannot be high without the h-index also being high. This indicates that the i10-index is more selective in nature

the i10 cannot be high without the h-index also being high. The collected data allows us to track the relationships between these two indexes across all authors in the dataset. The resulting log-log scale chart is presented in Fig. 2.

These two variables reveal a strong linear correlation ($R^2 = 0.892$) caused by their similar definitions. Log-log presentation makes 10 the critical point in terms of seriality. This means that below the value 10, there are single cases of distributed points, and that for the above values, the series of i10-index data can be observed. Seriality can be noted in the range of the h-index [10, 50] from the chart in Fig. 2. Quantitative proportions reveal the reverse state: for i10-index data below and over 10 values constitute approximately 77 percent and 23 percent of observations, respectively. The size and colour of markers allows the tracking of frequency dependences between indexes, revealing the most frequent pairs of variables for both values below 10.

Academic age of researchers – a pilot study

As previously mentioned, the variable 'earliest article year' consisted of two- or three-digit values in 87 cases, which have been excluded. The time elapsed from that year up to now is called the academic age taken into account during academic career studies (Milojević, 2012; Costas et al., 2015; Simoes & Crespo, 2020). The year 1960 was chosen as the cut-off year below which data was truncated. We assumed that the academic year of scholars presented on GS cannot exceed 60, taking into account (from autopsy) an experience of senior researchers with an electronic platform such as GS or RG. This assumption was justified, as only 266 records (1.3%) were excluded from the dataset. Only biological age data can confirm the correctness of this procedure, but access to such data is much more problematic (Kwiek & Roszka, 2022). Global-scale research requires substituting biological age with academic age or analysing their dependences. Kwiek and Roszka show that based on Scopus and administrative data, Polish researchers (*N* = 20,500) start their careers much later, in particular within no STEMM domains, than their colleagues from 'Western' countries. The histogram of the oldest article year is presented in Figure 3A. The distribution reveals three maxima at points: (I)1980, (II) 2002 and (III) 2012, which should be further studied in terms of seniority level. This need can be read particularly by binding a year with other variables, such as status or type. If we split the dataset into two status groups, we can see the differences in the distribution of year (Figure 3B).

It can be noted that young scholars with no more than a 10-year career (III) contributed most to IDUB universities, while non-IDUB scholars of academic age between 10 and 20 years are distributed equally (III and II groups, respectively). If Kwiek and Roszka (2022) found that there is a stronger correlation between biological and academic age in IDUBs (0.74) than in non-IDUBs (0.67), our observation can only give additional insight into the structure of the scientific community in Poland.

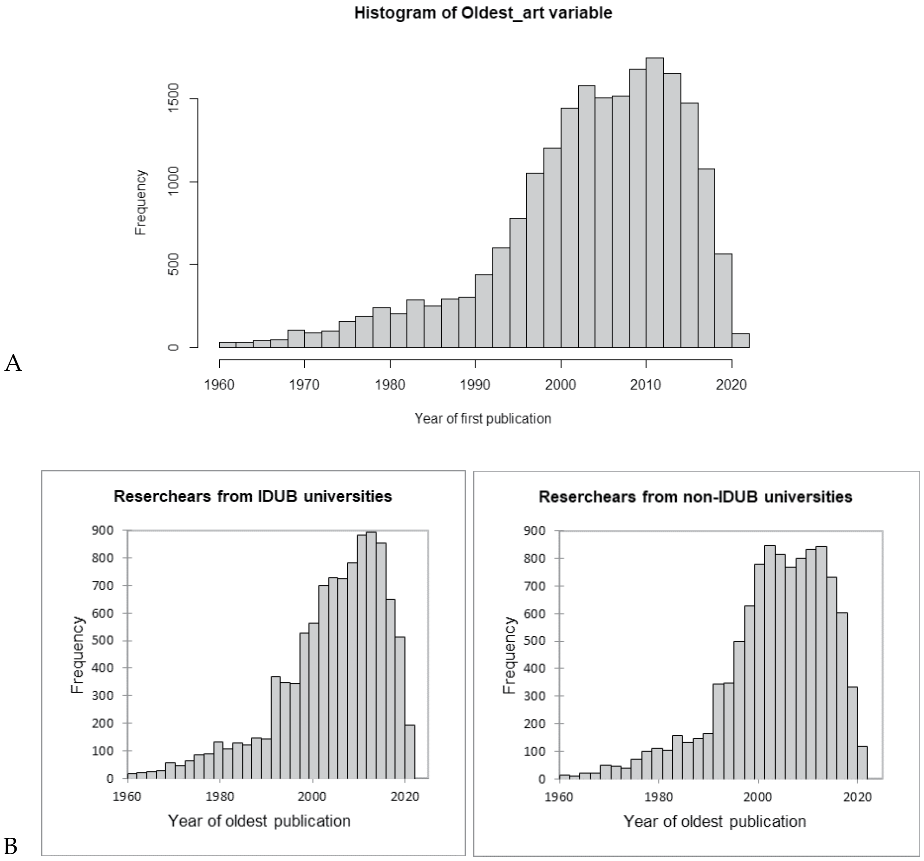


Fig. 3. The histogram of the earliest article year for the whole dataset (A) and split according to status variables (B)

4.3. LOGISTIC REGRESSION MODELS

The large-scale data derived from the network services is exposed to the randomness of errors to a greater or lesser extent. The results need to be more tangible (Williams, 2012); therefore, computing predicted or expected values and statistical hypotheses testing follow all processing phases. According to the initial aim of distinguishing two main groups in the working dataset, IDUB and non-IDUB, it can be determined whether or not the collected records belong to researchers from the school of excellence or not. A logistic regression model using the above-mentioned coding should answer such a question. Bornmann and Williams's research (2013), which concentrated on Leiden University rankings, used a logistic regression model for citation distribution. This approach was applied for institutional bibliometric evaluation. In logistic regression, the probability of an effect to be is non-linear (Ibidem). The response variable should be dichotomous in logistic regression-based models. In our case, university membership of the IDUB group is given a value of 1 for the response variable, which means the scholar belongs to an IDUB university, and zero if not.

Identifying differences in the impact of citation among universities can be performed by estimating a series of multivariate logistic regression models (Hosmer & Lemeshov, 2000; Bornmann & Williams, 2013). Thus, logistic regression models with tenfold cross-validation were created and evaluated. To select the best explanatory variables, a stepwise variable addition regression model was used. The criteria for including each variable were its statistical significance in the model. They were built based on the variables h-index, earliest article year, type, i10-index, as well as the newly created predictors below:

- **papers per year** containing an average number of publications yearly by counting from the year of the first publication to 2021 inclusively;
- **citations per paper** containing the average number of citations of the author's publication;
- **yearly citations per paper** containing the average number of citations for the publication in a year.

These new variables replaced the original ones "number of publications" and "citation count". The models were applied to the entire data set and served only for the selection of variables for the ultimate set. Finally, two models were selected to the evaluation: Model 1, based on predictors such as the h-index, the earliest article year, the type of university, and the number of papers per year; and Model 2, based on the h-index, the earliest article year, the type of university, the number of papers per year and the citation count. The latter was included in Model 2 despite a lack of statistical significance having been shown at the earlier stage because of potential improvement of the model's quality.

The cross-validation method was used to compare the models. The data was split into the training and test set ten times, such that in each sample, 10 percent of the observations remained in the test set, and each observation was in the test set only once. Then, a logistic regression model was created on the training set, and further, the classification on the test set was predicted and verified against actual values. In each model, for the elements of the test sets, the probability of being assigned to status = 1 (IDUB membership) was determined. Next, the probabilities of classification from the test sets were aggregated into a separate variable, and an ROC curve was generated (Figure 3A, B). Classifiers that give

ROC curves closer to the top-left corner indicate better performance. The closer the curve is to a 45-degree diagonal, the less accurate the model was performed. The AUC (area under the curve) is equal to 0.601 and 0.600 for Models 1 and 2, respectively. These values are the best for the tested models by using variables combined from GS data.

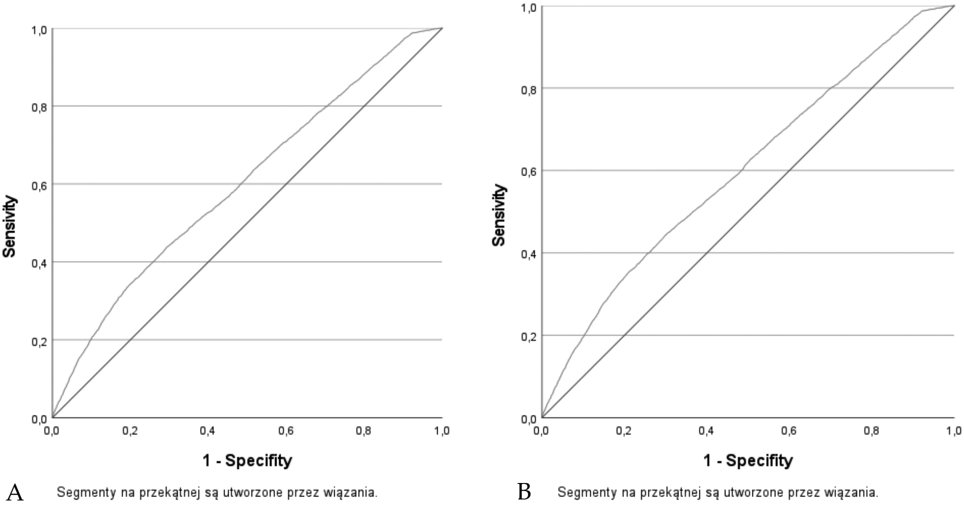


Fig. 4. The ROC curves for Model 1 (A) and Model 2(B). Sensitivity vs. specificity

To obtain validity and fitting coefficients between models, the logistic regression model was performed again for all data. Due to the similar values of the AUC of models, we considered Model 1, due to its smaller number of predictors. The quality assessment parameters obtained from the model are as follows: 55.3% of data was classified correctly (R^2 (Cox and Snell) = 0.038 and R^2(Nagelkerke) = 0.05) with the statistical significance ($p < 0.001$).

To improve the performance of the model, the optimal cut-off was determined using cross-validation for the ROC curve. Then, the mean value of these points was determined, obtaining the cut-off point for the ROC curve constructed for the entire sample. The value of 0.53423 was adopted as the cut-off point. Due to the fact that the point value is higher than 0.5, fewer observations will be in class 1, which implies an increase in specificity (measures of correctly classified zeros) and a decrease in sensitivity (measures of correctly classified ones). For the model used, the choice of the cut-off point increases the accuracy from 55% to 57% (Table 7). This means that the model predicts a correct fit more precisely than before the cut-off point was applied.

Table 7. Classification matrix of the model applying the cut-off approach

	Predicted classification	
Observed	**1**	**0**
1	4340	6095
0	2935	7673

Accuracy of model: 57%, sensitivity: 42%, specificity: 72%.

To determine performance differences between two universities, Bornmann and Williams (2013) proposed a logistic regression model to compare the predicted probabilities and predictive margins by averaging row values. Finally, it is necessary to study how the adjusted predictions for University 1 differ from those of University 2. According to the above-mentioned paper's approach in institutional evaluations, we can apply the average of predictive values returned by the logistic regression method.

By using the above-described model, the authors created a university ranking based on the average probability of assigning a scientist from this university to Status = 1. Three approaches to the calculations were used.

The first (4th column in Table 8) is based on averaging the probability of assigning a researcher to status = 1 if we group the whole dataset by university. The next two approaches use the concept of "average researcher" in a given university and the probability of being assigned to status = 1, understood as a record with average values of predictive variables calculated for scientists from the selected group. The third calculation (6th column, Table 8) is further modified by removing for each university the 5 percent of observations with the lowest number of citations and the 5 percent of observations with the highest number of citations.

Table 8. The ten best Universities in Poland according to the 2021 ranking of GS data and calculated from proposed models

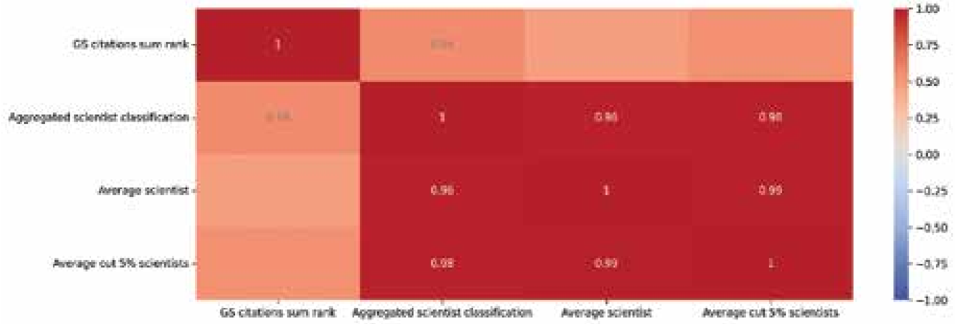
No	**University**	**GS citations sum rank**	**Aggregated scientist classification**	**Average scientist**	**Average cut 5% scientists**
1	Warsaw University (UW)	1	2	3	2
2	Gdańsk University of Technology(PG)	19	27	30	32
3	AGH University of Science and Technology (AGH)	4	21	22	25
4	Warsaw University of Technology (PW)	3	23	25	26
5	Adam Mickiewicz University (UAM)	10	1	1	1
6	Jagiellonian University (UJ)	2	5	8	9

7	Gdańsk Medical University (MUG)	24	37	36	38
8	Silesian University of Technology (PS)	14	28	28	27
9	Nicolaus Copernicus University (NCU)	5	6	5	4
10	Wrocław University (UWr)	9	3	2	3

To qualify an institution to join the set of excellence, the specified rankings were created. Table 8 presents the series of ranks for the top 10 Polish institutions based on input bibliometric data (sum of GS citations) and the used models.

As shown in Table 9, a statistically significant correlation is observed between total university citations and the variables used in the proposed models. Furthermore, the largest Spearman coefficients ($r = 0.996$, $p < 0.05$; and $r = 0.99$, $p < 0.05$) confirm the observations that the three approaches to calculating the probability of assigning the school to the IDUB category generate almost identical rankings.

Table 9. Spearman correlation matrix of variables which has been defined in Table 8



5. DISCUSSION

The selection of IDUB had to support the implementation of a new science policy. It should be noted that in the presented analyses we used data collected over a long period: even up to 60 years. However, during the 2013–2017-time window, as well as the overall, multifaceted potential and development plans decided about IDUB selection (see Ch.1.1). Furthermore, the GS data interface provides analyses of scientific output based on totals collected throughout the entire careers of individual citations and other measures; to obtain more detailed "hidden for user" information requires additional technical effort with an uncertain effect.

Official rankings of IDUB institutions illustrated in Table 1 show very high concordance by using Pearson correlation tests: $r = 0.946$, $p < 0.0001$. However, there is no observed statistically significant correlation between Scopus/Leiden rankings and those generated by the models in relation to IDUB items.

For a visual representation IDUBs positions in university rankings from Table 9, it is appropriate to use a parallel coordinates plot (Fig. 5). All three models stream to differentiate universities and polytechnics. The proportional dominance of overall university scholars within all IDUB records explains the favourable rankings of universities. Classification works better based on predictors built on the variables normalized to time unit and number of scholars employed in institutions.

A bigger and better representation of university researchers within all IDUB institutions in GS space can be combined with former recommendations made by an authority of a particular unit since 2011 (Kulczycki, 2011) as well as long-term, extensive information action organized by academic libraries (Lewandowski, 2014, 2017; Bogajczyk, 2019).

The logistic regression models illustrate which correlation effects are statistically significant and the direction of the effects (Bornmann & Williams, 2013). Adding new variables into models, we can track the strength of an effect, and this way control the model's performance. The first model of logistic regression used four explanatory variables: the h-index, the type, the oldest paper year and the number of articles yearly. The next model additionally included the number of citations. Both reveal almost the same performance. Measures based solely on citations do not necessarily indicate an IDUB university, whereas joint measures based on citations, publications count and period of publishing do indicate this.

It may turn out that the use of a logistic regression model is insufficient to understand the distribution of GS citations. The reason could be the uncertain, undetermined nature of internet-derived data.

GS offers citation patterns that are difficult or impossible to download from other national databases. Besides the individual level, valuable knowledge about data can come from the aggregation level. Aggregation-level analysis is facilitated by numerous possibilities of observation (scholars' profiles) grouping. Groups can be predefined by users arbitrarily, with an emphasis on deliberate goals. For example, grouping can be by gender, IDUB/non-IDUB schools or type of university. Thus, averaging the important bibliometric parameters according to groups can both provide insight into GS patterns and lead to incorrect conclusions. Therefore, we need to perform non-parametric and parametric statistical tests to confirm the differences in groups. These are statistically significant between medians and means for the two groups, IDUB schools and non-IDUB schools, if we consider the main scientometric measures. This observation is magnified by almost equal populations of these groups.

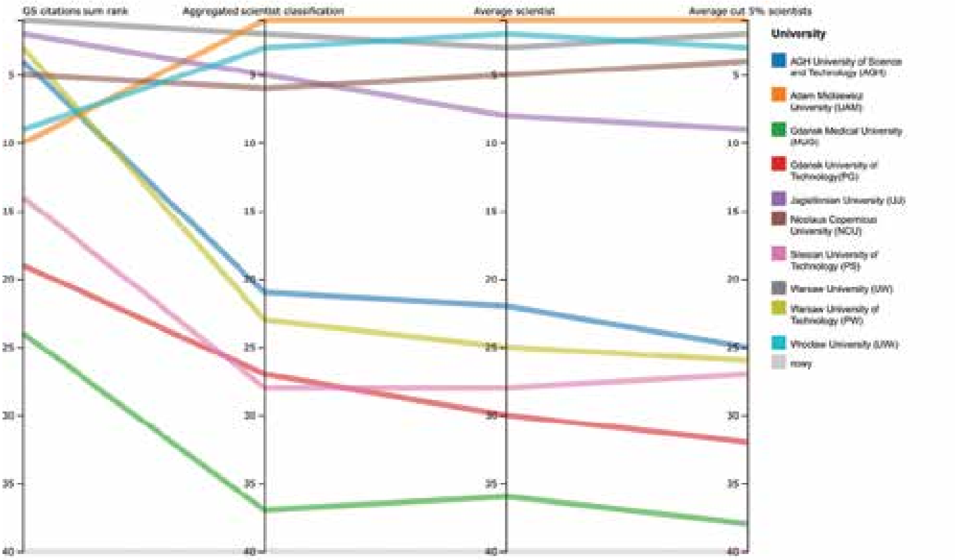


Fig. 5. IDUBs rankings according GS data and proposed models.

The distribution of the number of scholars according to type of university resembles the proportions in the official statistics, excluding medical schools, which are definitively under-represented. What we can say for certain is that medical scientists do not care about their visibility in GS space. Also, they do not manage the information in their accounts, which needs to be systematically updated. The empty lists of collaborators of many scholars with large numbers of citations demonstrate this neglect (Table 3).

6. CONCLUSION AND LIMITATIONS

GS data relating to Polish scientists has been processed and analysed in terms of citation distribution. Besides citations, other citation-based bibliometric indicators are taken into account to investigate any correlation of data with whether or not they belong to institutions of HE in Poland. Three years ago, ten universities of excellence (IDUBs) were officially selected within the framework of the new national science policy. The main aim of the current research was to examine whether scholars characterized by high scientometric measures were affiliated to IDUB schools. The hypothesis that individual scientists' indicators contribute to IDUBs ranking is confirmed within the scope of universities. For organizational reasons and according to their own specific policies, universities are represented best in GS space. This irregularity influences the modelling results, which favour universities over other types. The challenge may be to construct a model that is independent of the type of school. When truncating the extremes of the distribution in one of the models, the final rankings change little; therefore, the H3 hypothesis is confirmed.

Through a thorough data collection process, we found that GS data is highly dependent on scholars' willingness to create and curate profiles, in addition to their publishing activity. Furthermore, the study revealed a correlation between official institutional statistics and individual scholar data on GS, highlighting the usefulness of the platform in previewing researcher achievements.

According to the typology of the IDUB set, we collected data relating to different types of universities that are accessible on the GS platform (i.e., general universities, and medically and technically oriented universities). The most active in GS space are mainly university scholars (634.89 per institution), the least, medical university scholars (109.44 scholars per institution). Technical universities are located in the middle of this ranking (490.47 scholars per institution). A small representation of medical schools determines the highest averages of both citation count and the h-index. It should be noted that Polish universities overall are not satisfactorily represented on the GS platform. The percentual ratio of scholars existing on GS varies from a vague 6 percent to more than half (64%) (Table 2). Therefore, averaging the basic bibliometric parameters according to predefined groups using GS data can lead to incorrect conclusions. For example, the highest average citation count for medical schools of 1403.65 (Table 3) reached this value due to both a small number of these schools and scholars in the database.

However, by aggregating the data by university status originating from whether or not the university belongs to the IDUB group one can observe important relationships. The number of scholars is distributed almost equally between the subgroups IDUB and non-IDUB, despite the big variance of institution numbers of 10 and 34, respectively. All statistical averages per scholar and per institution of indexes are higher in the case of IDUBs (Table 5). Statistically significant differences between medians and means of basic scientometric measures (without favouring any) of these groups were found.

A parallel approach to the university metrics approach applied in the current study is to consider scholar achievements described by a particular row in the GS database. This way, it is possible to track essential correlations between variables, as well as build models to predict changes. By applying logistic regression models, the following essential observations can be made. Among the main scientometric indicators available on the GS platform, citations do not play a dominant role. Statistically significant relationships can be revealed using variables constructed from several essential characteristics of scholar overall activity across time: citations, oldest article year and number of publications. If we take all these variables into account, the first hypothesis will be confirmed, and we can answer the main research question in the affirmative. The low level of accuracy of logistic regression models (0.601 in the best case, with machine learning and cross-validation) might be explained by the low quality of GS data.

Another important observation is that GS data gives different representation for all types of universities in Poland. In particular, medical universities are under-represented in the GS environment, which interferes with the final confirmation of hypothesis H2 concerning the essential differences in the dataset according to the type of school.

At the scale of the entire GS data, dataset modelling allows only universities to fit into the first 10 (compare Table 1 and Table 9). Despite the common awareness that GS data is far from an ideal scientific database, it is worth listing its limitations. The disadvantages of GS data influencing decisions about emphasis on analysing groups instead of individuals are as follows:

- It is difficult to identify profiles clearly (fake or incorrect personal data profiles).
- It is difficult to verify if relevance to the institution is genuine or not, as retired scholars, postdocs, and visiting academics all leave a footprint on GS.
- It is difficult to identify student profiles.

It is unclear how up to date scientists keep their profiles and how often they check if the data therein is correct. There are likely more overinclusion errors, as people are more willing to accept too many citations rather than too few.

The following conclusions concerning data collection on the GS platform can be made:

1. GS does not cover the bibliometric characteristics of all Polish scholars and cannot be considered a representative bibliographic database.

2. The GS platform delivers no qualitative data, which means the traditional statistical methods using this data cannot explain all effect–causal correlations generated by scientific activity with high precision.

3. The GS citations count is not a sufficient characteristic of scholar activity in the context of statistical modelling and prediction. Row classification into IDUB and non-IDUB cannot be explained based only on the citation's variable; additional metrics such as the number of publications and academic age should also be considered.

4. Despite its relationships with both citations and number of (sorted) articles, the common scientometric measure h-index turned out to be an insufficient variable for modelling and further prediction.

5. Not all Polish scholars care about their visibility and information updating on the GS platform. In particular, this objection can be applied to researchers in the medical field.

Taking imperfections of the GS data into account, it is worth starting to talk about the GS platform that depends tightly on a scholar's activity in terms not only of the evaluation of researcher achievements but also their willingness to curate their own profiles. This finding can also indicate how to complement and repopulate existing scientometric data from the ground up by applying process automation such as machine learning algorithms. This experience can be used for the creation of scientometric indexes as close to scholar needs as possible.

We acknowledge that this study has limitations, and further research is necessary to validate the findings. The results have implications for policymakers, academic institutions, and individual researchers, and underscore the need for a more comprehensive approach to evaluating academic excellence. Current research may also be helpful in further comparative studies of the bibliographic measures of existing databases against those of GS.

Summarizing our efforts to obtain final patterns, we can identify several areas where more attention can be paid. The individual rankings of GS metrics need to perform more detailed research, placing particular emphasis on extracting other available characteristics such as example yearly changes of indexes or citation networks. GS data is worth studying in terms of relationships between the h-index and i10-index, as the latter is a metric giving unique knowledge. Gender correlations have great informative potential, and such research needs more attention on future authors' plans.

Preliminary research was performed into the distribution of scholar academic age. This needs more detailed and longitudinal studies, which, in combination with an analysis of citations' "long tails", would help identify ways of determining the seniority level based on non-qualitative, semi-structured data such as GS. It would then also be possible to test the last hypothesis.

FUNDING

The research leading to these results received funding from Narodowa Agencja Wymiany Akademickiej under Grant Agreement No PPN/BEK/2020/1/00345 (Recipient Adam Kola, PhD).

DECLARATIONS

The authors have no relevant financial or non-financial interests to disclose.

COMPLIANCE WITH ETHICAL STANDARDS

- Disclosure of potential conflicts of interest – NO
- Research involving Human Participants and/or Animals – NO
- Informed consent – NO

ANNEX

School	Full name
AGH	AGH University of Science and Technology
BUM	Medical University in Białystok
GUM	Gdańsk Medical University
LoUM	Medical University in Łódź
LUM	Lublin Medical University
PB	Białystok University of Technology
PBB	University of Bielsko-Biała
PCz	Częstochowa University of Technology
PG	Gdańsk University of Technology
PK	Tadeusz Kościuszko University of Technology
Pkie	Kielce University of Technology
PL	Lublin University of Technology
PLO	Łódź University of Technology
PO	Opole University of Technology
PP	Poznań University of Technology
PR	Rzeszów University of Technology
PRz	University of Technology
PS	Silesian University of Technology
PSz	University of Technology
PUM	Medical University in Poznań
PW	Warsaw University of Technology
PWr	Wrocław University of Science and Technology
SUM	Silesian Medical University
SzUM	Pomeranian Medical University
UAM	Adam Mickiewicz University
UB	University of Białystok
UG	University of Gdańsk
UJ	Jagiellonian University
UKie	Jan Kochanowski University
UKSW	Cardinal Stefan Wyszyński University
UKW	Casimir the Great University
UL	University of Łódź
UMCS	Maria Curie-Skłodowska University
UMK	Nicolaus Copernicus University
UO	Opole University

URz	University of Rzeszów
US	University of Silesia
USz	University of Szczecin
UW	University of Warsaw
UWM	University of Warmia and Mazury
UWr	University of Wrocław
UZ	University of Zielona Góra
WrUM	Wrocław Medical University
WUM	Warsaw Medical University

The 20 highest cited researchers on Google Scholar affiliated by Polish universities (uncertain scholars are shaded)

N	Name	Affiliation	IDUB or not	total_ cites	h_index	i10_index
1	Ponikowski P	WrUM	0	274848	168	822
2	Michal Tendera	SUM	0	210474	106	291
3	Michal Dwuznik	AGH	1	102117	125	235
4	Bartek Lipinski	UMK	1	93668	141	391
5	naomi breslau	Michigan State University	0	64450	124	251
6	Roman Topor-Madry	UJ	1	61252	69	131
7	Piotr Jaranowski	UBu	0	59264	90	210
8	Malgorzata Janik	PW	1	50941	116	320
9	Jan Lubinski	SzUM	0	42433	99	373
10	Jacek Namiesnik	PG	1	35623	88	623
11	Hania Szajewska	WUM	0	32535	90	250
12	Agnieszka Zagozdzinska	WP	1	31656	85	185
13	Tomasz Fiutowski	AGH	1	30655	85	312
14	Roman Slowinski	PP	0	29572	86	291
15	Oded Stark	UW	1	29414	54	126
16	Andrzej Skowron	UW	1	27112	64	272
17	Marian P. Kazmierkowski	Power Electronics and Drives	0	24702	50	115
18	Jacek Gronwald	SzUM	0	23668	70	234
19	Tomasz Guzik	UJ	1	23589	68	143
20	Francois Beguin	PP	0	23358	56	139

BIBLIOGRAPHY

Akaike, H. (21 December 1981), This Week's Citation Classic. *Current Contents Engineering, Technology, and Applied Sciences*, 12(51), 42 [Hirotogu Akaike comments on how he arrived at AIC].

Altman, D. G., & Bland, J. M. (1995). Statistics notes: the normal distribution. *Bmj*, 310(6975), 298.

Altman, D.G., & Bland, J. M. (1996). Detecting skewness from summary information. *Bmj*. 313(7066),1200.

Anders, Michael E., and Dennis P. Evans. "Comparison of PubMed and Google Scholar literature searches". Respiratory Care 55.5 (2010): 578-583.

Bar-Ilan, J. (2008). Which h-index? – a comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257–271. https://doi.org/10.1007/s11192-008-0216-y

Boeker, M., Vach, W., & Motschall, E. (2013). Google Scholar as a replacement for systematic literature searches: good relative recall and precision are not enough. *BMC medical research methodology*, 13(1), 1-12. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-131

Bogajczyk, M. (2019). *Naukowiec w sieci*. Warszawa: Biblioteka Uniwersytecka w Warszawie. https://www.buw.uw.edu.pl//wp-content/uploads/2019/09/Naukowiec-w-sieci_ebook-2.pdf

Bornmann, L., & Williams, R.A. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics*, 7(2), 562-574. https://doi.org/10.1016/j.joi.2013.02.005

Bornmann, L., Thor, A., Marx, W., & Schier, H. (2016). The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. *Journal of the Association for Information Science and Technology*, 67(11), 2778-2789.

Bramer, W.M., Giustini, D., Kramer, B.M. et al. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. *Systematic Reviews*, 2, 115 (2013). https://doi.org/10.1186/2046-4053-2-115

Buttliere, B., & Buder, J. (2017). Personalizing papers using Altmetrics: Comparing paper 'Quality'or 'Impact'to person 'Intelligence'or 'Personality'. *Scientometrics*, 111(1), 219-239

Chatterjee, A., Ghosh, A., Chakrabarti, B.K. (2016). Universality of Citation Distributions for Academic Institutions and Journals. *PLoS one*, 11(1): e0146762. https://doi.org/10.1371/journal.pone.0148863

Costas, R., Nane, GF., & Larivière, V. (2015). Is the year of first publication a good proxy of scholars' academic age? In A.A. Salah, Y. Tonta, A.A. Akdag Salah (Eds.). *Proceedings of the 15th international conference on scientometrics and informetrics* (pp. 988–998). Istanbul: Bogaziçi University Printhouse.

CWTS Leiden Ranking (2022). Retrieved April 4, 2022 from: https://www.leidenranking.com/information/indicators

D'Alessandro S. et al. (2020). Promote or Perish? A brief note on academic social networking sites and academic reputation. *Journal of Marketing Management,* 36, 5/6, p. 405-411.

Dorsch, I. (2017). Relative visibility of authors' publications in different information services. *Scientometrics,112*, 917-925.

Elliott, A. C., & Woodward, W. A. (2007). *Statistical analysis quick reference guidebook: With SPSS examples*. London: Sage.

Gusenbauer, M. (2019a). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1), 177-214 https://doi.org/10.1007/s11192-018-2958-51

Gusenbauer, M. (2019b). Suitable for Systematic Reviews and Meta-Analyses? The Capacity of 23 Academic Search Engines. *Academy of Management Annual Meeting Proceedings* 2019(1):12759 DOI: 10.5465/AMBPP.2019.12759abstract

Gusenbauer, M., & Haddaway, N. (2020), Which Academic Search Systems are Suitable for Systematic Reviews or Meta-Analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed and 26 other Resources. *Research Synthesis Methods,* 11(2), 181-217. https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1378

Harzing, A.W. (February 28, 2017). Google Scholar is a serious alternative to Web of Science. *Harzing.com. Research in International Management*. Retrieved July 27, 2021, from: https://harzing.com/blog/2017/02/google-scholar-is-a-serious-alternative-to-web-of-science

Harzing, AW., Alakangas, S. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics* 106, 787–804 (2016). https://doi.org/10.1007/s11192-015-1798-9

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Chichester, UK: John Wiley & Sons, Inc.

IDUB (2019). IDUB: Polska Inicjatywa Doskonałości – Uczelnie Badawcze. *Forum Akademickie,* 11; 12. https://prenumeruj.forumakademickie.pl/fa/2019/11/; https://prenumeruj.forumakademickie.pl/fa/2019/12/

Jacsó, P. (2008). Google Scholar revisited. *Online Information Review,* 32(1), 102-114. https://www.researchgate.net/publication/220207410_Google_Scholar_revisited

Jacsó, P. (2012). Google Scholar Metrics for Publications: The software and content features of a new open-access bibliometric service. *Online Information Review,* 36(4), 604-619.

Jensenius, F., Htun, M., Samuels, D., Singer, D., Lawrence, A., & Chwe, M. (2018). The Benefits and Pitfalls of Google Scholar. *PS: Political Science & Politics*, 51(4), 820-824.

Komunikat Ministra Nauki i Szkolnictwa Wyższego z dnia 11 maja 2018 r. o ustanowieniu przedsięwzięcia pod nazwą „Strategia Doskonałości – Uczelnia Badawcza" (2018). Retrieved September 23, 2021, from: https://www.gov.pl/web/edukacja-i-nauka/komunikat-ministra-nauki-i-szkolnictwa-wyzszego-z-dnia-11-maja-2018-r-o-ustanowieniu-przedsiewziecia-pod-nazwa-strategia-doskonalosci--uczelnia-badawcza

Kulczycki, E. (2011). Google Scholar Citations otwarte dla wszystkich! [online] Warsztat badacza, 17.11.2011. Retrieved March 21, 2022 from: https://ekulczycki.pl/warsztat_badacza/google-scholar-citations-otwarte-dla-wszystkich/

Kulczycki, E. (2013). Jak dodać prace do Google Scholar i zwiększyć liczbę cytowań oraz indeks Hirscha. Poradnik dla początkujących, Poznań: Stowarzyszenie EBIB. https://repozytorium.amu.edu.pl/bitstream/10593/4369/8/Jak_dodac_prace_do_Google_Scholar-v.1.1.pdf

Kwiek, M., Roszka, W. (2021). Gender-based homophily in research: A large-scale study of man-woman collaboration. *Journal of Informetrics*, 15(3), https://doi.org/10.1016/j.joi.2021.101171

Lewandowski, T. (2014). Google Scholar a repozytoria i biblioteki cyfrowe w Polsce. Otwarta Nauka, August 28, 2014. Retrieved March 23, 2022 from: https://otwartanauka.pl/analysis/case-studies/google-scholar-a-repozytoria-i-biblioteki-cyfrowe-w-polsce?showall=1&limitstart=

Lewandowski, T. (2017). Jak zwiększyć widoczność publikacji naukowych w Internecie z pomocą Google Scholar. Platforma Otwartej Nauki, November 21, 2017. Retrieved March 23, 2022 from: https://bg.uwb.edu.pl/DebataOA2017/materialy/lewandowski_oa_week_2017.pdf

López-Cózar, E.D., Robinson-García, N., & Torres-Salinas, D. (2012). Manipulating Google Scholar Citations and Google Scholar Metrics: simple, easy and tempting. *EC3 Working Papers* 6: 29 May, 2012. Retrieved June 6, 2021, from: https://arxiv.org/ftp/arxiv/papers/1212/1212.0638.pdf

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., López-Cózar, E.D. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations'COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126 (1), 871-906, https://doi.org/10.1007/s11192-020-03690-4 1 3

Methodology. CWTS Leiden Ranking (2014), Universiteit Leiden Centre for Science and Technology Studies. Retrieved April 4, 2022 from: https://www.leidenranking.com/Content/CWTS%20Leiden%20Ranking%202014.pdf

Milojević, S. (2012). How Are Academic Age, Productivity and Collaboration Related to Citing Behavior of Researchers? *PLoS one*, 7(11): e49176. https://doi.org/10.1371/journal.pone.0049176

Mingers, J., O'Hanley, J.R. & Okunola, M. (2017). Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics*, 113(1), 1627–1643. https://doi.org/10.1007/s11192-017-2532-6

Moed, H.F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, 10 (2), 533-551.

Nauka Polska (2021). Retrieved July 19, 2021, from: https://nauka-polska.pl/#/home/search?_k=ovh688

Norris, M., Oppernheim, Ch. (2007). Comparing alternatives to the *Web of Science* for coverage of the social sciences' literature. *Journal of Infometrics*, 1(2), 161-169.

Orcan, F. (2020). Parametric or non-parametric: Skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education,* 7(2), 255-265.

Pierwszy konkurs w programie „Inicjatywa doskonałości – uczelnia badawcza" (March 27, 2019). Konstytucja dla Nauki. Retrieved September 23, 2021 from:https://konstytucjadlanauki.gov.pl/pierwszy-konkurs-w-programie-inicjatywa-doskonalosci-uczelnia-badawcza

POL-on. Radon – raporty, analizy, dane (2021). Retrieved July 19, 2021, from: https://radon.nauka.gov.pl/raporty/Kadra2019

Prawo o szkolnictwie wyższym i nauce (2021). Obwieszczenie Marszałka Sejmu Rzeczypospolitej Polskiej z dnia 1 marca 2021 r. w sprawie ogłoszenia jednolitego tekstu ustawy – Prawo o szkolnictwie wyższym i nauce. *Dziennik Ustaw*, poz. 478, art. 388.

Pulikowski, A. (2015). Widoczność polskich publikacji naukowych w Internecie. *Zagadnienia Informacji Naukowej*, 53, 1 (105), 59-70.

Radicchi, F., Fortunato, S., Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS*, 105(45), 17268-17272. Retrieved April 3, 2022 from: www.pnas.orgcgidoi10.1073pnas.0806977105

Sauvayre, R. (2022). Types of Errors Hiding on Google Scholar Data. *Journal of Medical Internet Research*, 24 (5), pp. 1-13.

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association*, 95(4), 442-445. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2000776/

Simoes, N., & Crespo, N. (2020). A flexible approach for measuring author-level publishing performance. *Scientometrics*, 122(1), 331-355.

Świgoń, M., Głowacka, E., Kisilowska-Szurmińska, M.(2022). Academia.edu, ResearchGate, Google Scholar, Scopus i Publons (Web of Science) – szczegółowa analiza obecności reprezentantów nauk o komunikacji społecznej i mediach. *Media – Kultura – Komunikacja Społeczna*, 18, p. 83-101.

Teixeira da Silva, J. A. (2018). The Google Scholar h-index: useful but burdensome metric. *Scientometrics*, 117(1), 631-635. https://link.springer.com/article/10.1007%2Fs11192-018-2859-7

Teixeira da Silva, J. A. (2021). The i100-index, i1000-index and i10,000-index: expansion and fortification of the Google Scholar h-index for finer-scale citation descriptions and researcher classification. *Scientometrics*, 126 (4), 3667-3672. https://link.springer.com/article/10.1007%2Fs11192-020-03831-9

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PloS one*, *8*(5), e64841.

Thelwall, M., & Kousha, K. (2017). ResearchGate versus Google Scholar: Which finds more early citations?. *Scientometrics*, 112(2), 1125-1131. https://link.springer.com/article/10.1007%2Fs11192-017-2400-4

Thelwall, M., &Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, 8(4), 824-839. https://doi.org/10.1016/j.joi.2014.08.001

Tober, M. (2011). PubMed, ScienceDirect, Scopus or Google Scholar — Which is the best search engine for effective literature research in laser medicine?. *Medical Laser Application*, 26, 139-144.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., Van Eck, N.J., Van Leeuwen, T.N., Van Raan, A.F.J., Visser, M.S., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432.

Waltman, L., &Van Eck, N.J. (2013). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, 96(3), 699-716.

Wang, D., & Barabási, A. (2021). *Science of Science*. Boston. Publisher: Cambridge University Press.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, 12(2), 308-331.

Yang, K., & Meho, L.I. (2006), Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of American Society for Information Science and Technology*, 43: 1-15. https://doi.org/10.1002/meet.14504301185

Yu, G., Keirstead, J., Jefferis, G., Getzinger, G., Cimentada, J., Czapanskiy, M., & Makowski, D. (2021). *Analyse Citation Data from Google Scholar* [Package]. Retrieved April 4, 2022 from: https://cran.csail.mit.edu/web/packages/scholar/scholar.pdf

Zientek, L. R., Werner, J. M., Campuzano, M. V., & Nimon, K. (2018). The use of Google Scholar for research and research dissemination. *New Horizons in Adult Education and Human Resource Development*, 30(1), 39-46.

Artykuł w wersji poprawionej wpłynął do Redakcji 12 lutego 2024 r.

VESLAVA OSIŃSKA
Instytut Badań nad Informacją i Komunikacją
Uniwersytet Mikołaja Kopernika
e-mail: wieo@umk.pl
ORCID 0000-0002-1306-7832

BERNARDETA IWAŃSKA-CIEŚLIK
Instytut Komunikacji Społecznej i Mediów
Uniwersytet Kazimierza Wielkiego
e-mail: biwanska@ukw.edu.pl
ORCID 0000-0003-1841-6162

JAKUB WOJTASIK
Szkoła Doktorska Nauk Społecznych
Uniwersytet Mikołaja Kopernika
e-mail: jwojtasik@doktorant.umk.pl
ORCID 0000-0001-6157-5658

BRETT BUTTLIERE
Centrum Europejskich Studiów Regionalnych i Lokalnych(EUROREG)
Uniwersytet Warszawski
e-mail: brettbuttliere@gmail.com
ORCID 0000-0001-5025-0460

JOANNA KARŁOWSKA-PIK
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika
e-mail: joanka@mat.umk.pl
ORCID 0000-0001-9157-7355

ADAM KOLA
Wydział Humanistyczny
Uniwersyteckie Centrum Doskonałości "Interakcje – umysł, społeczeństwo, środowisko"
Instytut Badań Zaawansowanych, Uniwersytet Mikołaja Kopernika
Uniwersytet Amsterdamski, Holandia
e-mail: adamkola@umk.pl
ORCID 0000-0002-0584-6342

WKŁAD NAUKOWCÓW W RANKING IDUB (INICJATYWA DOSKONAŁOŚCI – UCZELNIA BADAWCZA). POLSCY BADACZE W GOOGLE SCHOLAR

SŁOWA KLUCZOWE: Google Scholar. Literatura naukowa. IDUB (Inicjatywa Doskonałości – Uczelnia Badawcza)

ABSTRAKT: **Teza/cel** – Google Scholar jest narzędziem szeroko wykorzystywanym nie tylko do wyszukiwania publikacji naukowych, ale także do uzyskiwania informacji na temat miar naukometrycznych dla poszczególnych badaczy. Autorzy artykułu weryfikują, czy na podstawie danych pozyskanych z Google Scholar użytkownicy z najwyższymi miarami zostaną zidentyfikowani jako badacze związani z najlepszymi uniwersytetami w Polsce (określanymi jako IDUB). **Metody badań** – Zastosowano modele krokowej regresji logistycznej z walidacją krzyżową, aby odszukać zmienne wpływające w znaczący sposób na poprawną klasyfikację automatyczną. **Wyniki i wnioski** – Jeśli chodzi o jakość przewidywania, najlepsze modele uzyskano przy użyciu następujących predyktorów: indeksu Hirscha (h-index), typu uniwersytetu, rocznej liczby publikacji oraz roku wydania pierwszej publikacji. Testy t-Studenta wykazały statystycznie znaczące różnice w średnich wartościach indeksu Hirscha, indeksu i10 oraz liczby publikacji (odpowiednio $p<0.001$, $p<0.001$ i $p=0.013$) pomiędzy naukowcami z 10 najlepszych uczelni w Polsce (IDUB) i badaczami z innych instytucji. Naukowcy charakteryzujący się wysokimi miarami naukometrycznymi są związani z uczelniami IDUB – związek ten obserwuje się w obrębie uniwersytetów, nie politechnik czy szkół medycznych. Swobodny i otwarty charakter Google Scholar sprawia, że pozyskiwane za jego pomocą dane są heterogeniczne i często niekompletne, co utrudnia ich automatyczne przetwarzanie i analizę. Utrudnienia te są szczególnie widoczne w przypadku agregacji danych. Pomimo tych ograniczeń pozyskane wyniki pozwalają jednak na zapanowanie nad szybkim przyrostem danych naukometrycznych i mogą prowadzić do powstania nowych miar oceny dorobku naukowego badaczy.