MERE 2021

META RESEARCH CONFERENCE

DECEMBER 1 & 3, 2021

Universitat Pompeu Fabra
Campus | Online

Davinia Hernández-Leo & J. Roberto Sánchez-Reina

Meta-Research

ICT Department, Universitat Pompeu Fabra
Barcelona January 2022

Editors

Davinia Hernández-Leo
Full Professor
Universitat Pompeu Fabra,
Barcelona
E-mail: davinia.hernandez-leo@upf.edu

J. Roberto Sánchez-Reina
Adjunt Professor
Universitat Pompeu Fabra
Barcelona
E-mail: roberto.sanchez@upf.edu

Universitat Pompeu Fabra, Barcelona
e-repository UPF, http://repositori.upf.edu/

Preface

Who investigates the work of scientists? Many organizations evaluate the work of scientists. The monitoring of the scientific work includes the institutional boards such as departments, faculties, and ethics committees that validate procedures to obtain grants, resources, or research licenses.

Yet, despite the existence of a large bureaucratic apparatus, 'researching on research' (the researchers' profession and their work) commonly falls outside the agenda of both disciplines and researchers. Except for peer review, and other formal practices, examining the scientific practice from an empirical perspective is not as attractive as other research objects.

How is science (and the scientists) facing the flourishing technological paradigms? Is scientific practice being affected by any ideological bias? How Open Science is contributing to more equal research and research careers? These are just a few questions raised by the presenters at the Meta Research Conference, MERE 2021.

The MERE Conference is an academic practice conducted every year by students of the Master's Programs in Sound and Music Computing, Intelligent and Interactive Systems, and Computational Biomedical Engineering of the Information and Communication Technology Department at Pompeu Fabra University.

The MERE 2021 is the result of an integrative assessment, the "Meta Research Project", that involves students of the Research Methods course to identify and evaluate the different elements that make up the scientific practice by researching research. For the MERE 2021 edition, the students have worked on a small piece of research to analyze: the principles and values of science, the role of open science, scientific communication and outreach, and ethics in research.

As part of the process, students have examined scientific research from an empirical perspective: they have systematized the literature, formulated hypotheses and questions, and validated them through observation or experimentation. Likewise, they have contributed to the evaluation process of their peers, and the academic debate during the conference.

Both procedures and results of MERE 2021 have been of scientific quality. A total of 18 manuscripts were submitted and defended to the Mere Conference. The present proceedings book compiles the papers generously shared by their contributors, who in exchange, seek to support scientific dialogue, discussion, and reflection.

We thank each student enrolled in the Research Method course and participants of the MERE Conference 2021. We also appreciate the kindness of authors who granted their work to be part of this group publication.

Prof. J. Roberto Sánchez-Reina
Prof. Davinia Hernández-Leo

CONTENT

Ethics in Research

1 Race And Ethnicity Bias: Are researchers aware of this problem?
Miriam Caravaca Rodríguez, Daniel Cañadas Gómez & Paula Chaves Hernández 1

2 Quantification of bias and its solutions in MICCAI challenges: towards standardization of bias mitigation
Laura Pérez, Andreu Pascuet, Marian Iglesias 16

3 The impact of GDPR: the researcher's perspective
Lois Riobó, Aina Albajar & Daniela Varela 39

Open Science

4 Equity in academic publishing: the impact of socioeconomic background on Open Science within Europe
Christina Zatse, Eva Encinas Crespo & Mariana Nakagawa 47

5 How Copyright Restrictions Affect Music Information Retrieval Research.
Dean Cochran, Betty Cortiñas-Lorenzo & Anna Barletta 56

6 Impact of Open Access to Datasets, A Case Study of Indian and Chinese Traditional Music.
Huicheng Zhang, Qingyuan Liu & Yuxi Qiao 68

7 Evaluation of countries based on their use of Open Science in medical research: a comparison with reference rankings
Eduard Alcobé, Miguel Silva & Aaron Verdaguer 85

8 An Evaluation of Readability Metrics for Scientific Writing
Ilse Meijer, Stephanie Rodríguez Osorio & Davide Locatelli 101

Race And Ethnicity Bias: Is machine learning a new form of discrimination?

Daniel Cañadas Gómez, Miriam Caravaca Rodríguez and Paula Chaves Hernández

Master in Computational Biomedical Engineering

{daniel.canadas01, miriam.caravaca01}@estudiant.upf.edu, paula.chaves@upf.edu

Abstract. Constant contributions to research in machine learning algorithms in the healthcare ambit are leading to improvements in health outcomes. These improvements can be, however, affected by the heterogeneity of the data used (diversity of patients) which, at the same time, depends on the difficulties to access to information of patients from different ethnic/racial groups (usually due to lack of the resources available). In this work we aimed to analyse recent publications from 2018 until now in Pubmed in the ambit of machine learning, to check if researchers are aware of this new form of discrimination. Our results indicate that there exists bias in a high percentage of the recent publications of machine learning in healthcare, and will still represent a problem of discrimination if new methodologies and data sets do not take it into account.

Keywords: machine learning, discrimination, ethics, artificial intelligence, racism

1. Introduction

Digital health is rapidly growing in most medical environments. By definition, digital health combines technologies such as mobile health, telemedicine or wearable devices to perform a wide range of medical utilities, for instance diagnosing, treating or supporting clinical decisions. In this ambit, machine learning (ML) is emerging as a highly promising tool to automatise some of these processes, by reducing time, costs and doctors' workload.

Despite the wide use of ML nowadays, and the obvious importance of their good performance in clinical environments, there still exists some bias in their functioning, often related to the quality of the data used to train these models. One of these sources of bias are related to the ethnicity of the patients. For example, in The Lancet Digital Health, Sarkar et al [1], analysed whether there exists or not an implicit bias in clinical severity scores, in COVID-19 patients in intensive care units. They found that there is a statistically significant difference across ethnicities, with a pattern of overprediction of mortality in

black people. This was also argued in A.Noseworthy et al. [2] and Turner Lee [3] who both express the necessity of reporting the performance of new machine learning health algorithms among diverse ethnic, racial, age and sex groups to avoid harming some of these groups and lead to new forms of discrimination.

This racial profiling can also be seen in biomedical research. It has been reported that the statistical black:white mortality from heart failure ratio is not true. [4] Based on these inaccurate statistics, some companies are developing "race-specific" therapy for heart failure like BiDil, who claimed that "observed racial disparities in mortality and therapeutic response rates in black patients may be due in part to ethnic differences in the underlying pathophysiology of heart failure" (NitroMed2001b). They were basing their claim in some underlying biological factor and didn't address any social or environmental factors. In research when "race" or ethnicity is used as an analytical variable, they rarely provide an explanation of how or why these variables are important.[5] This problem is acknowledged by some journals like *Nature Genetics.* Their solution is to oblige authors to "explain why they make use of particular ethnic groups or populations, and how classification was achieved."[6] This may help to avoid the differences being explained by ethnicity or race and to increase research to root out social injustice in medical practice.[7] So, in order to claim for a type of research that also examines racial/ethnic discrimination is important to review the procedures, regulations, policies and rules and what is the discriminatory impact [8]. The decision making process can affect a huge number of lives and also have life-and-death consequences [9][10].

As we have stated, despite that the constant advances in research and, in the last years, in machine learning are improving healthcare worldwide, there is still a certain intrinsic bias in all these procedures. Considering this fact into these practices, would help to personalize and adapt each treatment and procedure among the data related to the patient characteristics, thus equally improving patients' life quality. It is also possible that machine learning, when properly deployed, could help to resolve disparities in health care delivery if algorithms could be built to compensate for known biases or identify areas of needed research [3].

So the aim of this research is to explore if different machine learning researchers are facing this issue, by using heterogeneous data sets or adapted methodologies, and what further researchers must take into account to eliminate as much as possible this source of inequality. The objectives to accomplish are to analyse different ML articles in the ambit of healthcare and identify the proportion of researchers that are aware of this true problem, by analysing their data sets and methodologies.

2. Methodology

As previously mentioned, there is an increasing need of detecting extrinsic sources of racial bias in recent machine learning algorithms, which could slowly convert medicine into a discriminative science. To detect whether recent machine learning algorithms are conditioned by this type of bias, carrying out a meta-analysis of the recently published articles could be very useful. First of all, we wanted to compare how different common sources of bias are taken into consideration in machine learning research in the ambit of healthcare. To do this, we searched in Pubmed (https://pubmed.ncbi.nlm.nih.gov/) the keyword machine learning, and constrained the search to a series of keywords, to analyse whether the age, sex, ethnicity and race of the subjects are considered when creating the different datasets to later train the algorithms. The keywords used for each type of bias can be found in Table 1. The number of publications containing each of the keywords was recorded. Pubmed is a prestigious free resource supporting the search and retrieval of biomedical sciences publications, with the main objective of improving health, so this first step would give an overview of how the different types of bias could be influencing the results in ML publications.

Table 1. Keywords used to identify whether the different types of bias are considered or not.

Bias	Keyword
Age	Age OR ages
Sex	Sex OR male OR female
Ethnicity	Ethnics OR ethnicity
Race	Race OR racial

As a second step, we wanted to analyse more specifically the presence of ethnicity/racial bias in different Pubmed publications. To check this, 40 different articles in the ambit of machine learning were selected from the Pubmed dataset. This selection was completely randomized, so as to avoid adding external sources of bias. The data is distributed over 4 years from 2018 to 2021 with the analysis of 10 papers each year. These papers are referenced in Annex I.

The aim of this second part of our investigation, was to identify if researchers are aware of the probability of ethnic or racial bias when carrying out their projects and determine how this certain probability is faced. To do so, different aspects were identified in the studied articles.

First of all, we analysed the dataset used in each of the papers, and identified the location or locations in which the different data sets were created (country, hospital, whether the data comes from different parts of the world, etc.). Secondly, we checked if the data sets used in each project include patients from different ethnicities or races. In case they do, we checked if the data is balanced. An ethnicity/race balanced data set would include the same proportion of patients from different ethnicities/races, thus avoiding as much as possible this source of bias. Later, we analysed which of the articles took into account the possibility of bias by proposing alternative methods that would avoid this discrimination source, and which of them also validated their resulting algorithm in other countries (even if their data sets did not include different ethnicities/races, or the data was not balanced). Lastly, we identified if these articles, even when not considering the fact of ethnicity/racial bias, were taking into account other possible sources of bias, such as the age or sex bias. Those would give us an indicator about how concerned researchers are about racial bias in ML research.

Table 2. Questions used to identify how the probability of ethnic/racial bias is addressed in the different analysed articles.

	Questions
1	Where did they find the data?
2	Does the data include patients from different races/ethnicities?
3	If they do, is the data set balanced?
4	Does the methodology consider this fact?
5	Is the resulting algorithm validated in other countries?
6	Does it mention the probability of bias? (As a limitation)
7	Does it consider other types of bias? (Sex, age…)

3. Results

To start with, as mentioned we firstly compared how much the most common types of bias are taken into consideration in machine learning publications since 2011. Results can be found in Figure 1. In Fig.1.A, we can observe how the number of publications including keywords related to the different types of bias are growing in the recent years. As this type of increase could also be due to the exponential increase of ML publications, Fig.1.B and Fig.1.C show the normalized prevalence (number of articles containing the keyword divided by the total ML articles) of sex and age, and race and ethnicity related keywords, respectively. We can observe clear differences between Fig.1.B and Fig.1.C. While the age and sex seems to be the most considered sources of bias (both appear in around a 15% and 20% of the publications, respectively), ethnicity and race seem to be in most of the cases overlooked.

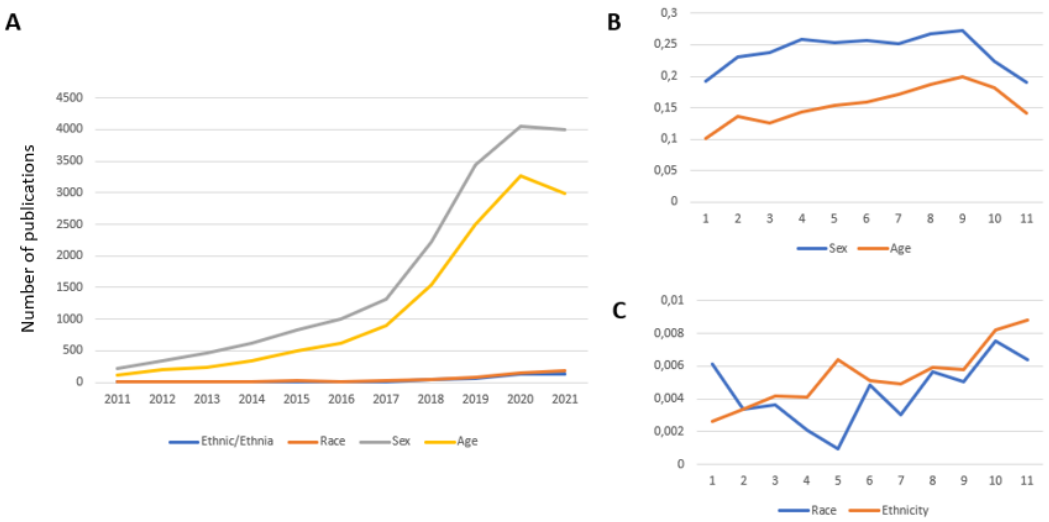


Figure 1. A. Presence of the keywords ethnic/ethnia, race, sex and age in the recently published articles in Pubmed, with respect to the year of publication. B. Normalized prevalence of sex and age keywords with respect to the number of publications in the ambit of machine learning each year. C. Normalized prevalence of race and ethnicity keywords with respect to the number of publications in machine learning.

As previously exposed, as a second step, we analysed 40 different machine learning publications in the ambit of machine learning, dated from 2018 until now. Different indicators were extracted out of these papers, and were used to identify if there is still an intrinsic ethnicity/racial bias in artificial intelligence research. The extracted indicators, in the form of a pie chart, can be observed in Figure 2.

First of all, we can see in Fig. 2.1. that sources of data in which machine learning algorithms were trained are mostly from the USA (47%) and Europe (42%). The next place in decreasing order is Asia (13%) with China being the majority (10%). Moreover, there weren't any articles that searched for data specifically in Africa, Australia or South-America. On the other hand, in Fig.2.4. we can see that the resulting algorithms, after being developed with the main source of data, are only validated in other countries in two articles out of the 40 we reviewed.

Regarding if the data includes different ethnicities or races, we can see in Fig. 2.2. that 10 out of the 40 articles did include them, but out of these articles the data isn’t balanced, meaning that, the percentages of each ethnicity or race are not equilibrated. White people were found in all the cases to be the major part of the population used in these studies, while black people represented only a 0-18% of the total data.

Finally, we also analysed if the papers did mention or take into consideration other types of biases apart from ethnicity or race bias. Fig. 2.3. shows that almost 78% of the papers take into account other types of biases, such as age or sex bias, by making equilibrated datasets (similar proportion of patients from different races/ethnicities) or mentioning the necessity of carrying further research to avoid possible biases. It can be seen then that there is an existing difficulty to collect data from patients from diverse race and ethnicities. The reason behind this fact, which was clearly identifiable in our meta-analysis could be the obstacles related to the lack of resources in certain parts of the world, as well as their laws and regulations. Although it is not always possible to deal with these obstacles, thus creating balanced data sets to train non-discriminant ML algorithms, it is vital that researchers inform in their writings about the bias probability that, in an unavoidable way, it exists in their results.

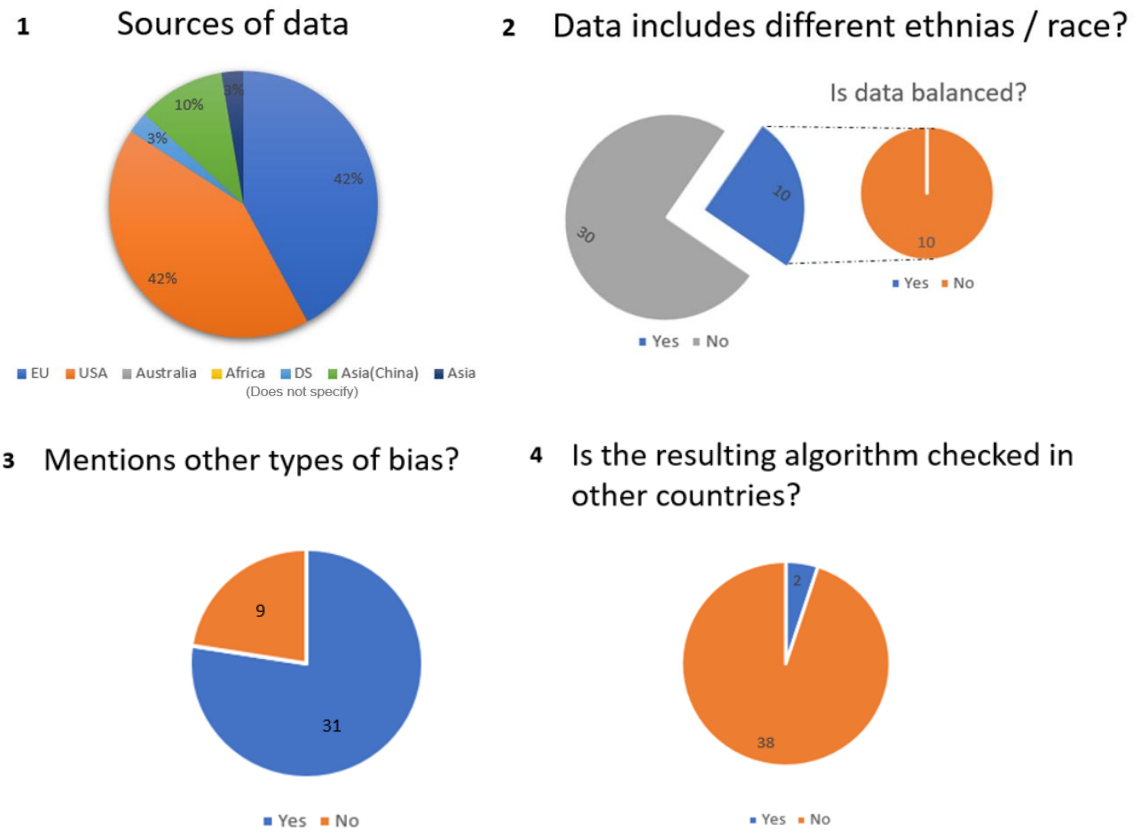


Figure 2. 1. Sources of data used to train the models for the 40 analysed research projects. 2. Percentage of papers considering in their studies patients from different ethnicities/races. Out of the projects including different ethnicities/races, percentage of papers in which the amount of patients from the different ethnicities are balanced.
3. Proportion of papers mentioning other types of bias, such as sex or age biases.
4. Proportion of research projects which checked their developed algorithm later in other countries, with patients from other races or ethnicities.

4. Conclusion

In this meta-research project we aimed at exploring if there exists an intrinsic bias in recent publications of machine learning in the ambit of healthcare. To do this, we firstly analysed the prevalence of different keywords associated with different types of bias in recent publications in the ambit of interest. Secondly, we selected 40 different articles in the same ambit and identified where did the data come from, if this data includes patients from different races or ethnicities, and if the algorithms are validated in other countries. We could conclude that race and ethnicity bias is still an existing discrimination problem that is not taken into account in most of the research projects, being rich countries from the European Union and USA the most common sources of data for these kinds of projects. We could also observe that researchers have difficulty, not only to extract high amounts of information from other countries to train their algorithms, but also to check later their developed models in all these countries. These problems could be due to a lack of resources in these parts of the world, cultural and political differences, among others.

We can conclude that further work has to be done in order to be able to access to more diverse datasets considering the different types of bias, and also to make the scientific community aware of the existing problem of discrimination in machine learning, for instance by mentioning the certain probability of bias that exists in ML projects using unbalanced data. As seen, the first step to end up with this source of inequality is to start using heterogeneous datasets including patients with different age, sex and racial features. For this reason, it is a major priority to embrace the international cooperation to acquire heterogeneous and balanced datasets for machine learning training algorithms. Only this way, we will avoid transforming healthcare into a discriminative science.

5. References

[1] Gumbsch, Thomas, & Borgwardt, Karsten. (2021). Ethnicity-based bias in clinical severity scores. The Lancet Digital Health, VOLUME 3, ISSUE 4, E209-E210. DOI:https://doi.org/10.1016/S2589-7500(21)00044-3

[2] Peter A. Noseworthy, MD, Zachi I. Attia, MSc, LaPrincess C. Brewer, MD, MPH, Sharonne N. Hayes, MD, Xiaoxi Yao, PhD, Suraj Kapa, MD, Paul A. Friedman, MD, Francisco Lopez-Jimenez, MD, MSc. (2020). Assessing and Mitigating Bias in Medical Artificial Intelligence. Circulation: Arrhythmia and Electrophysiology. Volume 13, No. 3. DOI: https://doi.org/10.1161/CIRCEP.119.007988

[3] Turner Lee, Nicol. (2018) Detecting racial bias in algorithms and machine learning. Journal of Information Communication and Ethics in Society 16(3). DOI:10.1108/JICES-06-2018-0056

[4] Kahn, J. (2003). Getting the numbers right: Statistical mischief and racial profiling in heart failure research. Perspectives in Biology and Medicine, 46(4), 473-83. Retrieved from https://www.proquest.com/scholarly-journals/getting-numbers-right-statistical-mischief-racial/docview/230776816/se-2?accountid=14708

[5] Lee, C. (2009). "Race" and "ethnicity" in biomedical research: How do scientists construct and explain differences in health? Social Science & Medicine, 68(6), 1183–1190. https://doi.org/10.1016/j.socscimed.2008.12.036

[6] Census, race and science. Nat Genet 2000;24:97-8.

[7] Schwartz, R. S., M.D. (2001). Editorial: Racial profiling in medical research. The New England Journal of Medicine, 344(18), 1392-1393. Retrieved from https://www.proquest.com/scholarly-journals/editorial-racial-profiling-medical-research/docview/223941672/se-2?accountid=14708

[8] Shavers, V. L., Klein, W. M. P., & Fagan, P. (2012). Research on race/ethnicity and health care discrimination: Where we are and where we need to go. *American Journal of Public Health*, *102*(5), 930–932. https://doi.org/10.2105/ajph.2012.300708

[9] Obermeyer, Ziad; Powers, Brian; Vogeli, Christine; Mullainathan, Sendhil (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. doi:10.1126/science.aax2342

[10] Bloo, G. J., Hesselink, G. J., Oron, A., Emond, E. J., Damen, J., Dekkers, W. J., Westert, G., Wolff, A. P., Calsbeek, H., & Wollersheim, H. C. (2014). Meta-analysis of operative mortality and complications in patients from minority ethnic groups. *The British journal of surgery*, *101*(11), 1341–1349. https://doi.org/10.1002/bjs.9609

Annex I: Analysed papers

2021

1. Blomberg SN, Christensen HC, Lippert F, Ersbøll AK, Torp-Petersen C, Sayre MR, Kudenchuk PJ, Folke F. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. JAMA Netw Open. 2021 Jan 4;4(1):e2032320. doi: 10.1001/jamanetworkopen.2020.32320. PMID: 33404620; PMCID: PMC7788469.
2. Qin H, Hu X, Zhang J, Dai H, He Y, Zhao Z, Yang J, Xu Z, Hu X, Chen Z. Machine-learning radiomics to predict early recurrence in perihilar cholangiocarcinoma after curative resection. Liver Int. 2021 Apr;41(4):837-850. doi: 10.1111/liv.14763. Epub 2020 Dec 25. PMID: 33306240.
3. Papp L, Spielvogel CP, Grubmüller B, Grahovac M, Krajnc D, Ecsedi B, Sareshgi RAM, Mohamad D, Hamboeck M, Rausch I, Mitterhauser M, Wadsak W, Haug AR, Kenner L, Mazal P, Susani M, Hartenbach S, Baltzer P, Helbich TH, Kramer G, Shariat SF, Beyer T, Hartenbach M, Hacker M. Supervised machine learning enables non-invasive lesion characterization in primary prostate cancer with [68Ga]Ga-PSMA-11 PET/MRI. Eur J Nucl Med Mol Imaging. 2021 Jun;48(6):1795-1805. doi: 10.1007/s00259-020-05140-y. Epub 2020 Dec 19. PMID: 33341915; PMCID: PMC8113201.
4. Ayers B, Sandholm T, Gosev I, Prasad S, Kilic A. Using machine learning to improve survival prediction after heart transplantation. J Card Surg. 2021 Nov;36(11):4113-4120. doi: 10.1111/jocs.15917. Epub 2021 Aug 19. PMID: 34414609.
5. Vodencarevic A, Tascilar K, Hartmann F, Reiser M, Hueber AJ, Haschka J, Bayat S, Meinderink T, Knitza J, Mendez L, Hagen M, Krönke G, Rech J, Manger B, Kleyer A, Zimmermann-Rittereiser M, Schett G, Simon D; RETRO study group. Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs. Arthritis Res Ther. 2021 Feb 27;23(1):67. doi: 10.1186/s13075-021-02439-5. PMID: 33640008; PMCID: PMC7913400.
6. Feng G, Zheng KI, Li YY, Rios RS, Zhu PW, Pan XY, Li G, Ma HL, Tang LJ, Byrne CD, Targher G, He N, Mi M, Chen YP, Zheng MH. Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD. J Hepatobiliary Pancreat Sci. 2021 Jul;28(7):593-603. doi: 10.1002/jhbp.972. Epub 2021 May 12. PMID: 33908180.
7. Machine Learning Consortium on behalf of the SPRINT Investigators. A Machine Learning Algorithm to Identify Patients at Risk of Unplanned Subsequent Surgery After Intramedullary Nailing for Tibial Shaft Fractures. J

Orthop Trauma. 2021 Oct 1;35(10):e381-e388. doi: 10.1097/BOT.0000000000002070. PMID: 34533505.
8. Xu X, Zhang J, Yang K, Wang Q, Chen X, Xu B. Prognostic prediction of hypertensive intracerebral hemorrhage using CT radiomics and machine learning. Brain Behav. 2021 May;11(5):e02085. doi: 10.1002/brb3.2085. Epub 2021 Feb 24. PMID: 33624945; PMCID: PMC8119849.
9. Varga TV, Liu J, Goldberg RB, Chen G, Dagogo-Jack S, Lorenzo C, Mather KJ, Pi-Sunyer X, Brunak S, Temprosa M; Diabetes Prevention Program Research Group. Predictive utilities of lipid traits, lipoprotein subfractions and other risk factors for incident diabetes: a machine learning approach in the Diabetes Prevention Program. BMJ Open Diabetes Res Care. 2021 Mar;9(1):e001953. doi: 10.1136/bmjdrc-2020-001953. PMID: 33789908; PMCID: PMC8016090.
10. Alexopoulos GS, Raue PJ, Banerjee S, Mauer E, Marino P, Soliman M, Kanellopoulos D, Solomonov N, Adeagbo A, Sirey JA, Hull TD, Kiosses DN, Areán PA. Modifiable predictors of suicidal ideation during psychotherapy for late-life major depression. A machine learning approach. Transl Psychiatry. 2021 Oct 18;11(1):536. doi: 10.1038/s41398-021-01656-5. PMID: 34663787; PMCID: PMC8523563.

2020

1. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, Schenk J, Terwindt LE, Hollmann MW, Vlaar AP, Veelo DP. Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. JAMA. 2020 Mar 17;323(11):1052-1060. doi: 10.1001/jama.2020.0592. PMID: 32065827; PMCID: PMC7078808.
2. Kagiyama N, Piccirilli M, Yanamala N, Shrestha S, Farjo PD, Casaclang-Verzosa G, Tarhuni WM, Nezarat N, Budoff MJ, Narula J, Sengupta PP. Machine Learning Assessment of Left Ventricular Diastolic Function Based on Electrocardiographic Features. J Am Coll Cardiol. 2020 Aug 25;76(8):930-941. doi: 10.1016/j.jacc.2020.06.061. PMID: 32819467.
3. Pavel AM, Rennie JM, de Vries LS, Blennow M, Foran A, Shah DK, Pressler RM, Kapellou O, Dempsey EM, Mathieson SR, Pavlidis E, van Huffelen AC, Livingstone V, Toet MC, Weeke LC, Finder M, Mitra S, Murray DM, Marnane WP, Boylan GB. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. Lancet Child Adolesc Health. 2020 Oct;4(10):740-749. doi: 10.1016/S2352-4642(20)30239-X. Epub 2020 Aug 27. PMID: 32861271; PMCID: PMC7492960.

4. Rozek DC, Andres WC, Smith NB, Leifker FR, Arne K, Jennings G, Dartnell N, Bryan CJ, Rudd MD. Using Machine Learning to Predict Suicide Attempts in Military Personnel. Psychiatry Res. 2020 Dec;294:113515. doi: 10.1016/j.psychres.2020.113515. Epub 2020 Oct 22. PMID: 33113452; PMCID: PMC7719604.
5. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, Askautrud HA, Pradhan M, Nesheim JA, Albregtsen F, Farstad IN, Domingo E, Church DN, Nesbakken A, Shepherd NA, Tomlinson I, Kerr R, Novelli M, Kerr DJ, Danielsen HE. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet. 2020 Feb 1;395(10221):350-360. doi: 10.1016/S0140-6736(19)32998-8. PMID: 32007170.
6. Albizu A, Fang R, Indahlastari A, O'Shea A, Stolte SE, See KB, Boutzoukas EM, Kraft JN, Nissim NR, Woods AJ. Machine learning and individual variability in electric field characteristics predict tDCS treatment response. Brain Stimul. 2020 Nov-Dec;13(6):1753-1764. doi: 10.1016/j.brs.2020.10.001. Epub 2020 Oct 10. PMID: 33049412; PMCID: PMC7731513.
7. Liu Y, Admon R, Mellem MS, Belleau EL, Kaiser RH, Clegg R, Beltzer M, Goer F, Vitaliano G, Ahammad P, Pizzagalli DA. Machine Learning Identifies Large-Scale Reward-Related Activity Modulated by Dopaminergic Enhancement in Major Depression. Biol Psychiatry Cogn Neurosci Neuroimaging. 2020 Feb;5(2):163-172. doi: 10.1016/j.bpsc.2019.10.002. Epub 2019 Oct 22. PMID: 31784354; PMCID: PMC7010544.
8. Sampedro-Gómez J, Dorado-Díaz PI, Vicente-Palacios V, Sánchez-Puente A, Jiménez-Navarro M, San Roman JA, Galindo-Villardón P, Sanchez PL, Fernández-Avilés F. Machine Learning to Predict Stent Restenosis Based on Daily Demographic, Clinical, and Angiographic Characteristics. Can J Cardiol. 2020 Oct;36(10):1624-1632. doi: 10.1016/j.cjca.2020.01.027. Epub 2020 Feb 7. PMID: 32311312.
9. Liu X, Hou Y, Wang X, Yu L, Wang X, Jiang L, Yang Z. Machine learning-based development and validation of a scoring system for progression-free survival in liver cancer. Hepatol Int. 2020 Jul;14(4):567-576. doi: 10.1007/s12072-020-10046-w. Epub 2020 Jun 18. PMID: 32556865.
10. Goldstein P, Ashar Y, Tesarz J, Kazgan M, Cetin B, Wager TD. Emerging Clinical Technology: Application of Machine Learning to Chronic Pain Assessments Based on Emotional Body Maps. Neurotherapeutics. 2020 Jul;17(3):774-783. doi: 10.1007/s13311-020-00886-7. PMID: 32767227; PMCID: PMC7609511.

2019

1. Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, Pouleur AC, Knappe D, Biering-Sørensen T, Kutyifa V, Moss A, Stein K, Solomon SD, Bijnens B. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. Eur J Heart Fail. 2019 Jan;21(1):74-85. doi: 10.1002/ejhf.1333. Epub 2018 Oct 17. PMID: 30328654.
2. Waljee AK, Wallace BI, Cohen-Mekelburg S, Liu Y, Liu B, Sauder K, Stidham RW, Zhu J, Higgins PDR. Development and Validation of Machine Learning Models in Prediction of Remission in Patients With Moderate to Severe Crohn Disease. JAMA Netw Open. 2019 May 3;2(5):e193721. doi: 10.1001/jamanetworkopen.2019.3721. Erratum in: JAMA Netw Open. 2019 Jun 5;2(6):e197386. PMID: 31074823; PMCID: PMC6512283.
3. Park A, Chute C, Rajpurkar P, Lou J, Ball RL, Shpanskaya K, Jabarkheel R, Kim LH, McKenna E, Tseng J, Ni J, Wishah F, Wittber F, Hong DS, Wilson TJ, Halabi S, Basu S, Patel BN, Lungren MP, Ng AY, Yeom KW. Deep Learning-Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. JAMA Netw Open. 2019 Jun 5;2(6):e195600. doi: 10.1001/jamanetworkopen.2019.5600. PMID: 31173130; PMCID: PMC6563570.
4. Lee JG, Ko J, Hae H, Kang SJ, Kang DY, Lee PH, Ahn JM, Park DW, Lee SW, Kim YH, Lee CW, Park SW, Park SJ. Intravascular ultrasound-based machine learning for predicting fractional flow reserve in intermediate coronary artery lesions. Atherosclerosis. 2020 Jan;292:171-177. doi: 10.1016/j.atherosclerosis.2019.10.022. Epub 2019 Nov 2. PMID: 31809986.
5. Cho H, Lee JG, Kang SJ, Kim WJ, Choi SY, Ko J, Min HS, Choi GH, Kang DY, Lee PH, Ahn JM, Park DW, Lee SW, Kim YH, Lee CW, Park SW, Park SJ. Angiography-Based Machine Learning for Predicting Fractional Flow Reserve in Intermediate Coronary Artery Lesions. J Am Heart Assoc. 2019 Feb 19;8(4):e011685. doi: 10.1161/JAHA.118.011685. PMID: 30764731; PMCID: PMC6405668.
6. Westcott A, Capaldi DPI, McCormack DG, Ward AD, Fenster A, Parraga G. Chronic Obstructive Pulmonary Disease: Thoracic CT Texture Analysis and Machine Learning to Predict Pulmonary Ventilation. Radiology. 2019 Dec;293(3):676-684. doi: 10.1148/radiol.2019190450. Epub 2019 Oct 22. PMID: 31638491.
7. Kim N, McCarthy DE, Loh WY, Cook JW, Piper ME, Schlam TR, Baker TB. Predictors of adherence to nicotine replacement therapy: Machine learning evidence that perceived need predicts medication use. Drug Alcohol Depend. 2019 Dec 1;205:107668. doi: 10.1016/j.drugalcdep.2019.107668. Epub 2019 Oct 25. PMID: 31707266; PMCID: PMC6931262.

8. Gadalla AAH, Friberg IM, Kift-Morgan A, Zhang J, Eberl M, Topley N, Weeks I, Cuff S, Wootton M, Gal M, Parekh G, Davis P, Gregory C, Hood K, Hughes K, Butler C, Francis NA. Identification of clinical and urine biomarkers for uncomplicated urinary tract infection using machine learning algorithms. Sci Rep. 2019 Dec 23;9(1):19694. doi: 10.1038/s41598-019-55523-x. PMID: 31873085; PMCID: PMC6928162.
9. Advanced Analytics Group of Pediatric Urology and ORC Personalized Medicine Group. Targeted Workup after Initial Febrile Urinary Tract Infection: Using a Novel Machine Learning Model to Identify Children Most Likely to Benefit from Voiding Cystourethrogram. J Urol. 2019 Jul;202(1):144-152. doi: 10.1097/JU.0000000000000186. Epub 2019 Jun 7. PMID: 30810465; PMCID: PMC7373365.
10. Reddy R, Resalat N, Wilson LM, Castle JR, El Youssef J, Jacobs PG. Prediction of Hypoglycemia During Aerobic Exercise in Adults With Type 1 Diabetes. J Diabetes Sci Technol. 2019 Sep;13(5):919-927. doi: 10.1177/1932296818823792. Epub 2019 Jan 17. PMID: 30650997; PMCID: PMC6955453.

2018

1. Lamping F, Jack T, Rübsamen N, Sasse M, Beerbaum P, Mikolajczyk RT, Boehne M, Karch A. Development and validation of a diagnostic model for early differentiation of sepsis and non-infectious SIRS in critically ill children - a data-driven approach using machine-learning algorithms. BMC Pediatr. 2018 Mar 15;18(1):112. doi: 10.1186/s12887-018-1082-2. PMID: 29544449; PMCID: PMC5853156.
2. Kalscheur MM, Kipp RT, Tattersall MC, Mei C, Buhr KA, DeMets DL, Field ME, Eckhardt LL, Page CD. Machine Learning Algorithm Predicts Cardiac Resynchronization Therapy Outcomes: Lessons From the COMPANION Trial. Circ Arrhythm Electrophysiol. 2018 Jan;11(1):e005499. doi: 10.1161/CIRCEP.117.005499. PMID: 29326129; PMCID: PMC5769699.
3. Schmidt-Erfurth U, Bogunovic H, Sadeghipour A, Schlegl T, Langs G, Gerendas BS, Osborne A, Waldstein SM. Machine Learning to Analyze the Prognostic Value of Current Imaging Biomarkers in Neovascular Age-Related Macular Degeneration. Ophthalmol Retina. 2018 Jan;2(1):24-30. doi: 10.1016/j.oret.2017.03.015. Epub 2017 May 31. PMID: 31047298.
4. Zilcha-Mano S, Roose SP, Brown PJ, Rutherford BR. A Machine Learning Approach to Identifying Placebo Responders in Late-Life Depression Trials. Am J Geriatr Psychiatry. 2018 Jun;26(6):669-677. doi: 10.1016/j.jagp.2018.01.001. Epub 2018 Jan 11. PMID: 29398354; PMCID: PMC5993576.

5. Schmidt-Erfurth U, Waldstein SM, Klimscha S, Sadeghipour A, Hu X, Gerendas BS, Osborne A, Bogunovic H. Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence. Invest Ophthalmol Vis Sci. 2018 Jul 2;59(8):3199-3208. doi: 10.1167/iovs.18-24106. PMID: 29971444.
6. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, Olsson EMG. Predicting Adherence to Internet-Delivered Psychotherapy for Symptoms of Depression and Anxiety After Myocardial Infarction: Machine Learning Insights From the U-CARE Heart Randomized Controlled Trial. J Med Internet Res. 2018 Oct 10;20(10):e10754. doi: 10.2196/10754. PMID: 30305255; PMCID: PMC6234350.
7. Basu S, Raghavan S, Wexler DJ, Berkowitz SA. Characteristics Associated With Decreased or Increased Mortality Risk From Glycemic Therapy Among Patients With Type 2 Diabetes and High Cardiovascular Risk: Machine Learning Analysis of the ACCORD Trial. Diabetes Care. 2018 Mar;41(3):604-612. doi: 10.2337/dc17-2252. Epub 2017 Dec 26. PMID: 29279299; PMCID: PMC5829969.
8. Koutsouleris N, Wobrock T, Guse B, Langguth B, Landgrebe M, Eichhammer P, Frank E, Cordes J, Wölwer W, Musso F, Winterer G, Gaebel W, Hajak G, Ohmann C, Verde PE, Rietschel M, Ahmed R, Honer WG, Dwyer D, Ghaseminejad F, Dechent P, Malchow B, Kreuzer PM, Poeppl TB, Schneider-Axmann T, Falkai P, Hasan A. Predicting Response to Repetitive Transcranial Magnetic Stimulation in Patients With Schizophrenia Using Structural Magnetic Resonance Imaging: A Multisite Machine Learning Analysis. Schizophr Bull. 2018 Aug 20;44(5):1021-1034. doi: 10.1093/schbul/sbx114. PMID: 28981875; PMCID: PMC6101524.
9. Nunez Lopez YO, Retnakaran R, Zinman B, Pratley RE, Seyhan AA. Predicting and understanding the response to short-term intensive insulin therapy in people with early type 2 diabetes. Mol Metab. 2019 Feb;20:63-78. doi: 10.1016/j.molmet.2018.11.003. Epub 2018 Nov 16. PMID: 30503831; PMCID: PMC6358589.
10. Dimai HP, Ljuhar R, Ljuhar D, Norman B, Nehrer S, Kurth A, Fahrleitner-Pammer A. Assessing the effects of long-term osteoporosis treatment by using conventional spine radiographs: results from a pilot study in a sub-cohort of a large randomized controlled trial. Skeletal Radiol. 2019 Jul;48(7):1023-1032. doi: 10.1007/s00256-018-3118-y. Epub 2018 Dec 1. PMID: 30506302; PMCID: PMC6525665.

Quantification of bias and its solutions in MICCAI challenges: towards standardization of bias mitigation

Maria Ángeles Iglesias Blanco, Andreu Pascuet Fontanet & Laura Pérez Sánchez

Master's in computational biomedical engineering
angeles.iglesias02@estudiant.upf.edu, andreu.pascuet01@estudiant.upf.edu, laura.perez32@estudiant.upf.edu

Abstract. Artificial intelligence has demonstrated efficiency and improvements regarding its use in healthcare. However, accumulating evidence demonstrates the impact of bias, which reflects social inequality on the performance of algorithms in healthcare. In this study, the consideration of bias in the Medical Image Computing and Computer Assisted Intervention grand challenge is analyzed to have an insight in the consideration of bias and its mitigation. A total of 349 papers were examined, from which 90 commented on the importance of bias. Although bias and solutions could be correlated into different fields, the different bias-mitigating solutions seem to have a place in standardization. Thereby, this study gives early insights towards standardization of solutions for AI bias and encourages future research for providing "fair" application of AI in healthcare.

Keywords: bias, healthcare, artificial intelligence, standardization, segmentation, machine learning

1 Introduction

Artificial Intelligence (AI) aims to imitate the proper functions of the human cognitive. Demonstrated efficiency and improvements in AI have caused an increase in its use in healthcare, powered by increasing availability of healthcare data and rapid progress of analytics techniques [1]. Indeed, image segmentation and machine learning (ML) are techniques that have widespread in medical fields. Segmentation has an essential role in computer-aided diagnosis systems attracting researchers to implement new medical image-processing algorithms [2]. Moreover, ML is an essential and effective tool for analysing highly complex medical data [3]. Because of the increase of AI applications in healthcare, competitions have been generated with the purpose of comparing methodologies given a particular task [4, 5]. Competitions are one of an increasing number of initiatives to achieve scientific progress by sharing data and comparing methods. The existence of repositories [6] allows researchers to use this data to train algorithms that they have created, with the aim of improving segmentation and cancer extraction between others. However, competitions serve more purposes than a comparative study of a range of algorithms on a common database. They also provide a snapshot of which methods are currently popular and in

the case of publishing the data and the algorithm used to perform a task, it can help for future projects, since they can be found in multiple state-of-the-art studies [5].

1.1 The Medical Image Computing and Computer Assisted Intervention (MICCAI) grand challenge

The performance of competitions has increased over the years, giving rise to the development of the Medical Image Computing and Computer Assisted Intervention (MICCAI) grand challenge, which was first organized in 2007 [7]. Since the first MICCAI competition, challenges have become an integral part of the MICCAI conferences. Furthermore, the datasets have become widely recognized as international benchmarking datasets and thus have a great influence on the research community and individual careers. Indeed, the impact of the challenges has increased so much that the development of these competitions is even a source of journal articles.

However, while the publication of papers in scientific journals and prestigious conferences, such as MICCAI, undergoes strict quality control including peer-reviewing, the design and organization for the participation in challenges do not [7, 8]. To perform these challenges, a training set is sent to the participants, who have some time to develop a methodology to accomplish the objective of the challenge, and then the algorithm is tested using a dataset that was not provided, with the aim of validating the performance in new data. In any case, those datasets have no explicit information in its quality and the consideration of bias when they were designed.

1.2 Bias in AI in healthcare

The increase of AI in healthcare has provoked recent awareness of the impacts of bias in AI algorithms and some doubts have been raised about the reliability of their use, especially because the algorithms may not be explainable in the same way that non-AI algorithms are [9]. And what is more, given the difficulty of explaining AI algorithms, reporting the bias affecting its outcomes is also a hard task. For a better understanding of the sources of these biases, Norori et al. [10] propose breaking down the different incomes of the bias into subcategories, so that they can be readily identified. At a great scale, the sources of the bias can be divided into: data-driven, which means that collected information is not representative for the human population as a whole, but only for a specific group of people or scenario; algorithmic, for which bias appears once the algorithm is trained, so they are harder to identify; and human, in which the problem is not about the used methodology, but on its subjective use by the researchers and its interpretation, mainly related with socioeconomical biases such us ethnicity, gender and sexual orientation [10].

Accumulating evidence demonstrates the impact of bias that reflects social inequality on the performance of algorithms in healthcare. Given their intended placement within healthcare decision making more broadly, they require attention to adequately quantify the impact of bias and reduce its potential to exacerbate inequalities [11]. Several studies with AI-based diagnostics have already shown social

or statistical bias in the reported results in the real-world application, such as the Framingham Study risk factors which has been used for decades to predict risk of suffering a cardiovascular disease [12-15].

To solve that, several efforts are being put into establishing a framework for reducing the bias in AI healthcare, or at least being able to detect and report it. On the one hand, different processes are emerging for mitigating the impact of bias on current predictions and reducing its impact over time, such as data regularization or outcomes analysis [9, 16]. On the other hand, major regulatory institutions such as the European Commission are regularly updating guidelines like the Coordinated Plan on Artificial Intelligence for avoiding bias in AI-healthcare [17]. However, all the proposed guidelines and processes for diminishing the presence of biases in AI-healthcare are still at its infancy, as shown by the fact that they need to be regularly updated to be effective. Therefore, there is a need for having a better-defined regulatory framework from which bias can be advised, corrected, and avoided.

Having highlighted the importance of the presence of bias in a model when translated to the real clinical application, this meta-research aims to assess how bias is being reported and addressed in AI and healthcare in a transversal way, across different tasks and applications. More concretely, it is sought to make a critical analysis of past MICCAI challenges in the segmentation and ML fields, since segmentation tasks are the most common ones in MICCAI challenges and ML has proven to be an essential and effective tool for AI implementations [18, 19]. By quantifying how many of these projects are considering bias when developing a model in the healthcare field, it is intended to clear out which areas have raised more awareness about the impact of bias, and to point out which are the currently proposed solutions, establishing if they are enough and determining which paths should be followed to have safe and ethical use of AI-based tools.

More information on how these questions are addressed can be found in the following sections. Firstly, in the Methods section, a brief and concise description of the process followed for acquiring the data and analyzing it is given. Then, Results are presented in forms of graphs and are contextualized in the Conclusions of the paper, in which the different inferences lead to the proposal of new guidelines on how to address bias in AI-healthcare.

2 Research methodology

For the development of this study, an exhaustive search for bias in the MICCAI conference proceedings of 2020 and 2021 was performed. The volumes analyzed were those regarding segmentation tasks and ML applications which are of great interest in the clinical field as discussed in the Introduction [2, 3]. Table 1 shows the number of papers included in the MICCAI challenges of 2020 and 2021 for the corresponding volumes. Once the information was collected, results were analyzed in Microsoft Excel 365[1] and ideas were proposed in order to solve the bias problem.

[1] Software available in https://www.microsoft.com/es-es/microsoft-365/excel

Table 1. MICCAI conference proceedings volumes analyzed. The year, volume and number of papers for each volume can be found.

MICCAI year	Volume	Number of papers
2020	I: Machine learning	81
	IV: Segmentation	79
2021	I: Image Segmentation	69
	II: Machine Learning 1	60
	III: Machine Learning 2	60

The analysis of the MICCAI proceedings was based on an advanced OR search looking for the presence of synonyms that indicated the existence of bias in the challenges. The word selection leaded to the consideration of the following concepts: bias, inequality, prone, race, racial, gender, sex, economy and economic; based on principal synonyms for "bias" and specific concepts that are correlated with bias in the literature. Once works considering bias were identified, they were further analyzed to be fit in different classifications for the drawing of conclusions. On the one hand, papers were subdivided in three categories depending on the type of bias they presented (data-driven, algorithmic and human), according to the classification proposed by Norori et al. [10] seen at the Introduction; and on the task they were performing, based on the finding of tasks through the reading of the papers. On the other hand, a further classification of the considered papers was performed by only considering those that applied some type of correction for reducing the bias. From those, a classification between the ones corrected by refining the data or modifying the algorithm was also performed. For an overview of the process, see Figure 1.

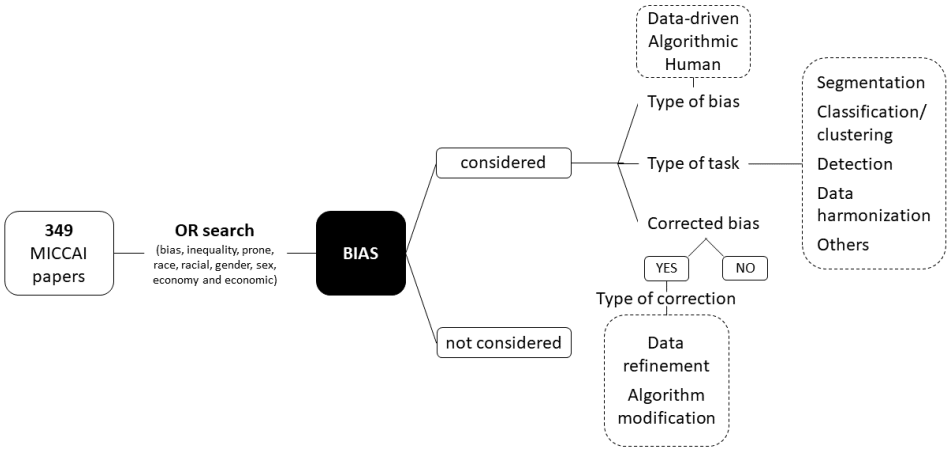


Figure 1. Workflow followed for the study of the bias in the MICCAI challenges of 2020 and 2021. Firstly, 349 papers were analyzed using an OR search to find the existence of bias. Secondly, according to the presence of synonyms of bias, they were classified into bias considered or not considered. Finally, the ones that commented on the importance of bias were organized based on the type of bias, the type of task and the type of correction if done.

3 Results

349 papers from the Segmentation and ML Volumes of the 2020 and 2021 MICCAI challenges were considered, of which 90 considered bias based on the OR search described in the Research methodology section (see Appendix section for the list of found papers). Those considering bias could be further classified into the type of bias that they presented, having 44 papers with algorithmic bias, 41 papers with data-driven bias and 5 papers with human bias (see Figure 2).

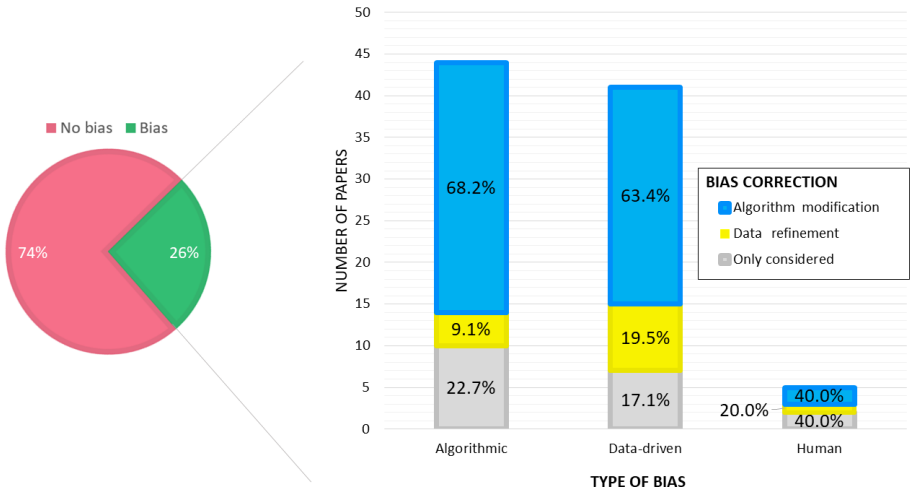


Figure 2. Bias quantification, classification and correction in Segmentation and ML tasks in MICCAI's 2020 and 2021 challenges. The left pie chart shows the percentage of papers that mention or not bias. From the ones that mention bias, the right stacked chart prompts the type of bias in the bars, and how it was corrected is shown in the stacked regions of each bar.

79% of the papers considering bias presented a type of correction, either an algorithm modification (64%) or data refinement (15%). The resting 21%, mentioned the presence of bias, but did not propose any methodology to correct it (see Figure 3). Figure 2 also shows percentages of how bias was corrected depending on the type of bias that was present in each paper, revealing non-significant differences between the general distribution of types of correction and the distribution in each of the types of bias, except for the human bias, for which there are not enough examples to drive a significant correlation.

After reading the papers mentioning bias, 5 general AI tasks were observed in the works from the ML volumes, more than a half corresponding to segmentation task. Next, classification and data harmonization tasks are present in a similar percentage (13,3% and 12,2%, respectively), followed by detection and other tasks (6,7% and 5,6%, respectively).

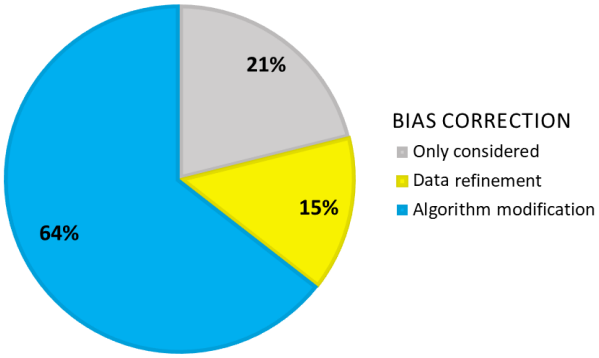


Figure 3. Distribution of types of bias correction. From the papers that commented on the existence of bias, percentages of those corrected by data refinement, algorithm modification or that have only considered the existence of bias, but not corrected it.

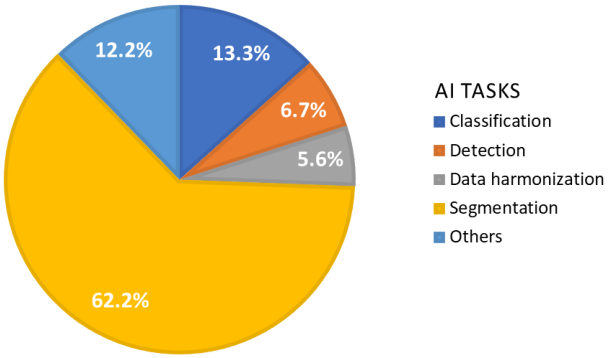


Figure 4. Distribution of AI tasks. From the papers that commented on the existence of bias in the ML volumes, percentages of those that performed a classification, detection, data harmonization, segmentation, or another task. Inside other tasks there are included: regression, registration, calibration, data augmentation, clinical reports and evaluation of performance.

4 Conclusions

The main objective of the present study is to determine if bias is really being considered when developing a model intended for a healthcare application due to its important concerns in safety and efficiency and, consequently, to its translation into the real world. The AI pipelines mainly contain three possible points of intervention to mitigate unwanted bias: the training data, the learning procedure, and the output

predictions, and these are associated with three corresponding classes of bias mitigation strategy: pre-processing (data correction), in-processing (algorithm correction), and post-processing (output prediction correction) [20]. By taking that into account, this meta-research seeks to quantify the types of bias in a reference challenge such as MICCAI, and to provide examples of how they are addressed to better define a strategy for mitigating bias.

Only 90 (26%) of the 349 papers considered in this study commented on the importance of bias, as shown in Figure 2. This prompts the lack of consideration of bias in studies, even when talking on peer-reviewed articles. But what is more, from the 26% that commented on the importance of bias, a total of 19 (21%) papers only considered the presence of bias (see Figure 3), without trying to overcome this issue. This is translated into a direct spreading of bias inside the scientific community, as long as publications without mentioning bias may be considered for future works that can exacerbate the bias, and those that mention the bias but do not correct it may encourage other researchers on following the same path. In any case, there are still 79% of papers mentioning bias that include a solution, and that are further analyzed to draw conclusions.

The majority of the correction of bias (64%) was faced by an algorithmic approach (see Figure 3), such as presenting a new method or architecture of the network [21]; defining a new loss metric [22] or using some concreted kernels to avoid bias [23]. A 15% presented a solution implicated with the data processing correction (see Figure 3), for instance the refinement of the data (increase of the resolution or the quality) [24]; an increase of the data labeled by experts [25]; the augmentation of data [26]; dealing with imbalanced data [27]; a preprocessing to eliminate some kind of the bias before using any algorithm [28]; a selection of data to avoid using the data that presents low quality [29]… among a long list of other possible methods depending on the data. However, none of the papers presented an example of a post-processing solution, being the reason why it was not taken into consideration for the classification of bias correction groups.

According to the classification of bias, the great majority was defined as algorithmic (49%) and data driven (46%). Algorithmic biases include, among others, the inductive bias associated to the structure of neural networks [30, 31, 32] or the type of cost function in which the optimization algorithm is based on [33, 34]. Regarding the data-driven bias, the most repeated ones are those regarding the imbalance of the data [30, 35, 36, 37]; not enough realistic data; or bad representations of the real distribution of a studied problem [38, 39]. However, human bias was almost unconsidered: only 5 out of the 349 papers talked about it and, of them, only 3 dealt with it (see Figure 2). Although the models presented in these challenges are not thought to be immediately applied to real world problems, there should be some space to consider possible biases, and human ones should be further highlighted. The assessment of the performance of a model should surpass the train and test usual methodology if “fairness” wants to be achieved in this kind of algorithms. More concretely, different studies have presented problems by not considering the differences between socioeconomic status [40, 41]. These studies assumed that medical expenditures equate to healthcare needs, leading to black patients being less likely to be identified by the algorithm as candidates for potentially beneficial care programs than white patients, who had the same number of chronic

illnesses. In one of these cases [40], remedying this disparity would increase the percentage of black patients receiving additional help from 17.7 to 46.5%. AI models are starting to be deployed in the real world, hence it is essential that the benefits of AI are shared equitably according to race, gender and other demographic characteristics, and so efforts to ensure the fairness of deployed models have generated much interest [20, 42].

Besides segmentation tasks being the most common ones in MICCAI challenges, it has been seen that most ML applications that consider and correct bias are intended also for segmentation (62%) (see Figure 4). This may be due to the importance of preventing bias in segmentation because of the close relation it has with clinical problems. Developing automatic, accurate, and robust medical image segmentation methods has been one of the principal problems in medical imaging as it is essential for computer-aided diagnosis and image-guided surgery systems. Segmentation of organs or lesions from a medical scan helps clinicians make an accurate diagnosis, plan the surgical procedure, and propose treatment strategies. In Figure 4, it is also highlighted other ML applications that considered bias: classification (13%), detection (7%) and data harmonization (6%). Other applications found are registration, calibration, regression, data augmentation, clinical reports, evaluation of the performance.

Overall, there is a huge combination of processes that can mitigate the impact of bias on current predictions and reduce its impact over time, some of them already implemented in the reviewed articles. Among them, analyzing the training set to ensure its suitability for the project; reviewing the inputs to test their correctness; evaluating the predictions so that they are grounded in reasonable feature values; incorporate new outcomes into future predictions to overcome historical bias; encourage to transparency and comprehensiveness when writing protocols for trials that evaluate AI interventions and when reporting its results… [7, 43]. And besides recommendations and guidelines, there have been also proposed different approaches and standards to measure the risk of bias of a study, i.e., PROBAST (Prediction model Risk Of Bias ASsessment Tool), a tool for assessing the risk of bias (ROB) and applicability of diagnostic and prognostic prediction model studies; or surveys of the different methodologies that can be applied to the most common problems in AI, i.e., a recent one concerning the ML field provided by Mehrabi et Al. [16, 44].

Despite all the proposed tools, the main hurdle to achieve fairness in AI is to find a standard way to quantify and address bias due to the amount of heterogeneity of the algorithms and the source of bias that the training data can contain if used, among others. In line with this work, human bias seems to be the hardest to overcome not only because it is less identified, but also because of its nature, since it affects the basis of the model design, and hence can be present in other non-AI fields. However, this problem must not be put aside, and if bias cannot be solved, at least some kind of measurement and comments regarding the possibility of bias should be provided, to be able to evaluate the concerns of its applicability.

After an exhaustive revision of the state-of-the-art it seems that the basis towards fairness in AI are being established. From the papers which addressed bias and the state-of-the-art review, it can be seen that there is a vast source of bias and possible solutions to it. The most prominent source but also solution of bias is the algorithmic one, which can be interpreted as a bottleneck for standardization. However, this can

also be taken as an advantage for treating biased data. In healthcare, data has a limited accessibility due to the data protection of the patient, heterogeneity in procedures, different machine acquisition, etc. These factors are mainly unmodifiable, forcing researchers to apply algorithm modifications to deal with this, so the observed quantifications match with the expected. In any case, efforts towards generating high-quality healthcare databases that can be easily shared should not be substituted by algorithmic solutions, as they are proving to be a real breakthrough towards fairness in AI [10].

All in all, this study provides a quantification of in which scale bias is being considered and how it is usually addressed depending on the causation source. Apparently, this problem is not being conceived with the relevance that it should, since from addressing bias, it depends on fairness in AI. In any case, the methodology is based only on a few volumes of the MICCAI challenges, so further revisions of other volumes, challenges and even keywords for the OR search should be considered to have a broader view of the state of bias in AI healthcare applications. From commented solutions for bias, standardizations to mitigate AI bias should go towards strategically deploying AI and carefully selecting underlying data. However, commitment by the researchers to understand the sources of bias is also needed, and meta-research projects such as the presented one should go hand in hand with regulatory committees that provide guidelines on how to manage research without spreading bias implications. If everything above could be accomplished, addressing bias could allow AI to reach its fullest potential by helping to improve diagnosis and prediction while protecting patients.

References

1. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. In Stroke and Vascular Neurology (Vol. 2, Issue 4). https://doi.org/10.1136/svn-2017-000101
2. Guo, Y., Guo, Y., & Ashour, A. (2019). Neutrosophic set in medical image analysis. San Diego: Elsevier Science & Technology
3. Garg, A., & Mago, V. (2021). Role of machine learning in medical research: A survey. Computer Science Review, 40, 100370. https://doi.org/10.1016/j.cosrev.2021.100370
4. Van Ginneken, B., Heimann, T., & Styner, M.A (2007). 3d segmentation in the clinic: A grand challenge. Workshop on 3D Segmentation in the Clinic, MICCAI 2007, pages 7–15
5. Styner, M., Lee, J., Chin, B., Chin, M. S., Commowick, O., Tran, H., Jewells & V., Warfield, S. (2008). 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. Workshop on 3D Segmentation in the Clinic, MICCAI 2008
6. Dua, D. & Graff, C. (2019). UCI Machine Learning Repository. Retrieved November 2021, from http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.
7. Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., & Full, P. et al. (2018). How to Exploit Weaknesses in Biomedical Challenge Design and Organization. Medical Image Computing And Computer Assisted Intervention – MICCAI 2018, 388-395. https://doi.org/10.1007/978-3-030-00937-3_45

8. MICCAI. (n.d.). Springer. Retrieved November 2021, from https://www.springer.com/gp/computer-science/lncs/societies-and-lncs/miccai/734372
9. Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in AI. The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019. https://doi.org/10.1145/3308560.3317590
10. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns (New York, N.Y.), 2(10), 100347. https://doi.org/10.1016/j.patter.2021.100347
11. McCradden, M., Joshi, S., Anderson, J., Mazwi, M., Goldenberg, A., & Zlotnik Shaul, R. (2020). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. Journal Of The American Medical Informatics Association, 27(12), 2024-2027. https://doi.org/10.1093/jamia/ocaa085
12. Canto, J. G., Goldberg, R. J., Hand, M. M., Bonow, R. O., Sopko, G., Pepine, C. J., & Long, T. (2007). Symptom presentation of women with acute coronary syndromes: Myth vs reality. In Archives of Internal Medicine (Vol. 167, Issue 22). https://doi.org/10.1001/archinte.167.22.2405
13. Gijsberts, C. M., Groenewegen, K. A., Hoefer, I. E., Eijkemans, M. J. C., Asselbergs, F. W., Anderson, T. J., Britton, A. R., Dekker, J. M., Engström, G., Evans, G. W., de Graaf, J., Grobbee, D. E., Hedblad, B., Holewijn, S., Ikeda, A., Kitagawa, K., Kitamura, A., de Kleijn, D. P. V., Lonn, E. M., … den Ruijter, H. M. (2015). Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. PLoS ONE, 10(7). https://doi.org/10.1371/journal.pone.0132321
14. McCarthy, A. M., Bristol, M., Domchek, S. M., Groeneveld, P. W., Kim, Y., Motanya, U. N., Shea, J. A., & Armstrong, K. (2016). Health care segregation, physician recommendation, and racial disparities in BRCA1/2 testing among women with breast cancer. Journal of Clinical Oncology, 34(22). https://doi.org/10.1200/JCO.2015.66.0019
15. Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. In JAMA - Journal of the American Medical Association (Vol. 322, Issue 24). https://doi.org/10.1001/jama.2019.18058
16. Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1). https://doi.org/10.7326/M18-1376
17. Bahrke, J., & Manoury, C. (2021, April 26). New rules for Artificial Intelligence – Q&As. European Comission. Accessed on 2nd November 2021. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683
18. Guo, Y., & Ashour, A. S. (2019). Neutrosophic Set in Medical Image Analysis (1st ed.). Academic Press. https://doi.org/10.1016/C2018-0-01943-X
19. Garg, A., & Mago, V. (2021). Role of machine learning in medical research: A survey. Computer Science Review, 40, 100370. https://doi.org/10.1016/j.cosrev.2021.100370
20. Anton, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., & King, A.P. (2021). Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. (2021) MICCAI 2021 conference.
21. Zhang Y., Liu H., Hu Q. (2021) TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_2
22. Shen Y., Jia X., Meng M.QH. (2021) HRENet: A Hard Region Enhancement Network for Polyp Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in

Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_53

23. Tang Y. et al. (2021) Pancreas CT Segmentation by Predictive Phenotyping. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_3
24. Chen Z. et al. (2021) A Novel Hybrid Convolutional Neural Network for Accurate Organ Segmentation in 3D Head and Neck CT Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_54
25. Wang K. et al. (2021) Tripled-Uncertainty Guided Mean Teacher Model for Semi-supervised Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_42
26. Yang J., Zhang Y., Liang Y., Zhang Y., He L., He Z. (2021) TumorCP: A Simple but Effective Object-Level Data Augmentation for Tumor Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_55
27. Li H. et al. (2021) Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_4
28. Yu Z., Zhai Y., Han X., Peng T., Zhang XY. (2021) MouseGAN: GAN-Based Multiple MRI Modalities Synthesis and Segmentation for Mouse Brain Structures. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_42
29. Xu Z. et al. (2021) Noisy Labels are Treasure: Mean-Teacher-Assisted Confident Learning for Hepatic Vessel Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_1
30. Daza L., Pérez J.C., Arbeláez P. (2021) Towards Robust General Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_1
31. You D., Liu F., Ge S., Xie X., Zhang J., Wu X. (2021) AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_7
32. Gu R., Zhang J., Huang R., Lei W., Wang G., Zhang S. (2021) Domain Composition and Attention for Unseen-Domain Generalizable Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_23
33. Van Tulder G., Tong Y., Marchiori E. (2021) Multi-view Analysis of Unregistered Medical Images Using Cross-View Transformers. In: de Bruijne M. et al. (eds) Medical Image

Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_10

34. Marrakchi Y., Makansi O., Brox T. (2021) Fighting Class Imbalance with Contrastive Learning. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_44
35. Khakzar A. et al. (2021) Towards Semantic Interpretation of Thoracic Disease and COVID-19 Diagnosis Models. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_47
36. Liu J., Guo X., Yuan Y. (2021) Prototypical Interaction Graph for Unsupervised Domain Adaptation in Surgical Instrument Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_26
37. Larrazabal A.J., Martínez C., Dolz J., Ferrante E. (2021) Orthogonal Ensemble Networks for Biomedical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_56
38. Liu Q., Yang H., Dou Q., Heng PA. (2021) Federated Semi-supervised Medical Image Classification via Inter-client Relation Matching. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_31
39. Henn T. et al. (2021) A Principled Approach to Failure Analysis and Model Repairment: Demonstration in Medical Imaging. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_48
40. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464). https://doi.org/10.1126/science.aax2342
41. Wiens, J., Price, W. N., & Sjoding, M. W. (2020). Diagnosing bias in data-driven algorithms for healthcare. In Nature Medicine (Vol. 26, Issue 1). https://doi.org/10.1038/s41591-019-0726-6
42. Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. In JAMA - Journal of the American Medical Association (Vol. 322, Issue 24, pp. 2377–2378). American Medical Association. https://doi.org/10.1001/jama.2019.18058
43. Wynants, L., Smits, L. J. M., & van Calster, B. (2020). Demystifying AI in healthcare. In The BMJ (Vol. 370). https://doi.org/10.1136/bmj.m3505
44. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. In ACM Computing Surveys (Vol. 54, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3457607

5 Appendix

The revised papers, corresponding to those which contained some of the key words, are listed below.

5.1 MICCAI 2020

VOLUME I: Machine learning

I. Marzahl C. et al. (2020) Are Fast Labeling Methods Reliable? A Case Study of Computer-Aided Expert Annotations on Microscopy Slides. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_3

II. Quan L., Li Y., Chen X., Zhang N. (2020) An Effective Data Refinement Approach for Upper Gastrointestinal Anatomy Recognition. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_5

III. Shi W., Xu K., Song M., Fan L., Jiang T. (2020) Constrain Latent Space for Schizophrenia Classification via Dual Space Mapping Net. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_9

IV. Liu X., Tsaftaris S.A. (2020) Have You Forgotten? A Method to Assess if Machine Learning Models Have Forgotten Data. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_10

V. Sheikh R., Schultz T. (2020) Feature Preserving Smoothing Provides Simple and Effective Data Augmentation for Medical Image Segmentation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_12

VI. Li Z., Zhong C., Wang R., Zheng WS. (2020) Continual Learning of New Diseases with Dual Distillation and Ensemble Strategy. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_17

VII. Zhang L. et al. (2020) Learning to Segment When Experts Disagree. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_18

VIII. Bône A., Vernhet P., Colliot O., Durrleman S. (2020) Learning Joint Shape and Appearance Representations with Metamorphic Auto-Encoders. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_20

IX. Qin Y. et al. (2020) Learning Bronchiole-Sensitive Airway Segmentation CNNs by Feature Recalibration and Attention Distillation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_22

X. Varsavsky T., Orbes-Arteaga M., Sudre C.H., Graham M.S., Nachev P., Cardoso M.J. (2020) Test-Time Unsupervised Domain Adaptation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_42

XI. Bozorgtabar B., Mahapatra D., Vray G., Thiran JP. (2020) SALAD: Self-supervised Aggregation Learning for Anomaly Detection on X-Rays. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_46

XII. Bateson M., Kervadec H., Dolz J., Lombaert H., Ben Ayed I. (2020) Source-Relaxed Domain Adaptation for Image Segmentation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_48

XIII. Unnikrishnan B., Nguyen C.M., Balaram S., Foo C.S., Krishnaswamy P. (2020) Semi-supervised Classification of Diagnostic Radiographs with NoTeacher: A Teacher that is Not Mean. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_61

XIV. Yang H., Shan C., Kolen A.F., de With P.H.N. (2020) Deep Q-Network-Driven Catheter Segmentation in 3D US by Hybrid Constrained Semi-supervised Learning and Dual-UNet. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_63

XV. Chen C. et al. (2020) Realistic Adversarial Data Augmentation for MR Image Segmentation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_65

XVI. Venturini L., Papageorghiou A.T., Noble J.A., Namburete A.I.L. (2020) Uncertainty Estimates as Data Selection Criteria to Boost Omni-Supervised Learning. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in

Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_67

XVII. Gonzalez Duque V., Al Chanti D., Crouzier M., Nordez A., Lacourpaille L., Mateus D. (2020) Spatio-Temporal Consistency and Negative Label Transfer for 3D Freehand US Segmentation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_69

XVIII. Hemsley M. et al. (2020) Deep Generative Model for Synthetic-CT Generation with Uncertainty Predictions. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_81

XIX. Zhou Y., Chen H., Lin H., Heng PA. (2020) Deep Semi-supervised Knowledge Distillation for Overlapping Cervical Cell Instance Segmentation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_51

XX. Ye J. et al. (2020) Synthetic Sample Selection via Reinforcement Learning. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_6

XXI. Zheng H. et al. (2020) Cartilage Segmentation in High-Resolution 3D Micro-CT Images via Uncertainty-Guided Self-training with Very Sparse Annotation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_78

XXII. Lee H., Jeong WK. (2020) Scribble2Label: Scribble-Supervised Cell Segmentation via Self-generating Pseudo-Labels with Consistency. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_2

XXIII. Wei D., Cao S., Ma K., Zheng Y. (2020) Learning and Exploiting Interclass Visual Correlations for Medical Image Classification. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_11

XXIV. Cui H., Xu Y., Li W., Wang L., Duh H. (2020) Collaborative Learning of Cross-channel Clinical Attention for Radiotherapy-Related Esophageal Fistula Prediction from CT. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham. https://doi.org/10.1007/978-3-030-59710-8_21

VOLUME IV: Segmentation

I. Wolleb J., Sandkühler R., Cattin P.C. (2020) DeScarGAN: Disease-Specific Anomaly Detection with Weak Supervision. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_2

II. Boot T., Irshad H. (2020) Diagnostic Assessment of Deep Learning Algorithms for Detection and Segmentation of Lesion in Mammographic Images. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_6

III. Lin JY., Chang YC., Hsu W.H. (2020) Efficient and Phase-Aware Video Super-Resolution for Cardiac MRI. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_7

IV. Dong G. et al. (2020) TexNet: Texture Loss Based Network for Gastric Antrum Segmentation in Ultrasound. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_14

V. La Rosa F. et al. (2020) Automated Detection of Cortical Lesions in Multiple Sclerosis Patients with 7T MRI. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_57

VI. Song Y. et al. (2020) Shape Mask Generator: Learning to Refine Shape Priors for Segmenting Overlapping Cervical Cytoplasms. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_62

VII. Dai C. et al. (2020) Suggestive Annotation of Brain Tumour Images with Gradient-Guided Sampling. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_16

VIII. Mo S. et al. (2020) Multimodal Priors Guided Segmentation of Liver Lesions in MRI Using Mutual Information Based Graph Co-Attention Networks. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_42

IX. Jiang X. et al. (2020) Multi-phase and Multi-level Selective Feature Fusion for Automated Pancreas Segmentation from CT Images. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_45

X. Sun J., Darbehani F., Zaidi M., Wang B. (2020) SAUNet: Shape Attentive U-Net for Interpretable Medical Image Segmentation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_77

XI. Shirokikh B. et al. (2020) Universal Loss Reweighting to Balance Lesion Size Inequality in 3D Medical Image Segmentation. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. Springer, Cham. https://doi.org/10.1007/978-3-030-59719-1_51

5.2 MICCAI 2021

VOLUME I: Segmentation

I. Xu Z. et al. (2021) Noisy Labels are Treasure: Mean-Teacher-Assisted Confident Learning for Hepatic Vessel Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_1

II. Zhang Y., Liu H., Hu Q. (2021) TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_2

III. Tang Y. et al. (2021) Pancreas CT Segmentation by Predictive Phenotyping. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_3

IV. Valanarasu J.M.J., Oza P., Hacihaliloglu I., Patel V.M. (2021) Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_4

V. Wang J., Wei L., Wang L., Zhou Q., Zhu L., Qin J. (2021) Boundary-Aware Transformers for Skin Lesion Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_20

VI. Song Y., Yu L., Lei B., Choi KS., Qin J. (2021) Selective Learning from External Data for CT Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_40

VII. Yu Z., Zhai Y., Han X., Peng T., Zhang XY. (2021) MouseGAN: GAN-Based Multiple MRI Modalities Synthesis and Segmentation for Mouse Brain Structures. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_42

VIII. Shen Y., Jia X., Meng M.QH. (2021) HRENet: A Hard Region Enhancement Network for Polyp Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_53

IX. Chen Z. et al. (2021) A Novel Hybrid Convolutional Neural Network for Accurate Organ Segmentation in 3D Head and Neck CT Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_54

X. Yang J., Zhang Y., Liang Y., Zhang Y., He L., He Z. (2021) TumorCP: A Simple but Effective Object-Level Data Augmentation for Tumor Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_55

XI. Yang J., Gu S., Wei D., Pfister H., Ni B. (2021) RibSeg Dataset and Strong Point Cloud Baselines for Rib Segmentation from CT Scans. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_58

XII. Popordanoska T., Bertels J., Vandermeulen D., Maes F., Blaschko M.B. (2021) On the Relationship Between Calibrated Predictors and Unbiased Volume Estimation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_64

XIII. Wei J., Hu Y., Zhang R., Li Z., Zhou S.K., Cui S. (2021) Shallow Attention Network for Polyp Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_66

XIV. Yang J., Hu X., Chen C., Tsai C. (2021) A Topological-Attention ConvLSTM Network and Its Application to EM Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_21

XV. Li L., Lian S., Luo Z., Li S., Wang B., Li S. (2021) Learning Consistency- and Discrepancy-Context for 2D Organ Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.

MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_25

XVI. Ma Q., Zu C., Wu X., Zhou J., Wang Y. (2021) Coarse-To-Fine Segmentation of Organs at Risk in Nasopharyngeal Carcinoma Radiotherapy. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_34

XVII. Zhou Y. et al. (2021) Learning to Address Intra-segment Misclassification in Retinal Imaging. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_46

XVIII. Nguyen T., Hua BS., Le N. (2021) 3D-UCaps: 3D Capsules Unet for Volumetric Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_52

XIX. Shi J., Wu J. (2021) Distilling Effective Supervision for Robust Medical Image Segmentation with Noisy Labels. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_63

XX. Ou Y. et al. (2021) LambdaUNet: 2.5D Stroke Lesion Segmentation of Diffusion-Weighted MR Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_69

VOLUME II: Machine learning part 1

I. Li H. et al. (2021) Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_4

II. Dufumier B. et al. (2021) Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_6

III. Ouyang J. et al. (2021) Self-supervised Longitudinal Neighbourhood Embedding. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_8

IV. Zhao F. et al. (2021) Learning 4D Infant Cortical Surface Atlas with Unsupervised Spherical Networks. In: de Bruijne M. et al. (eds) Medical Image

Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_25

V. Kamraoui R.A., Ta VT., Papadakis N., Compaire F., Manjon J.V., Coupé P. (2021) POPCORN: Progressive Pseudo-Labeling with Consistency Regularization and Neighboring. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_35

VI. Wang R., Wu Y., Chen H., Wang L., Meng D. (2021) Neighbor Matching for Semi-supervised Learning. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_41

VII. Wang K. et al. (2021) Tripled-Uncertainty Guided Mean Teacher Model for Semi-supervised Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_42

VIII. Wang S. et al. (2021) CPNet: Cycle Prototype Network for Weakly-Supervised 3D Renal Compartments Segmentation on CT Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_55

IX. Sambaturu B., Gupta A., Jawahar C.V., Arora C. (2021) Efficient and Generic Interactive Segmentation Framework to Correct Mispredictions During Clinical Evaluation of Medical Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_58

X. Liu X., Thermos S., O'Neil A., Tsaftaris S.A. (2021) Semi-supervised Meta-learning with Disentanglement for Domain-Generalised Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_29

XI. Yeung PH., Namburete A.I.L., Xie W. (2021) Sli2Vol: Annotate a 3D Volume from a Single Slice with Self-supervised Learning. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_7

XII. Huang Y., Lin L., Cheng P., Lyu J., Tang X. (2021) Lesion-Based Contrastive Learning for Diabetic Retinopathy Grading from Fundus Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_11

XIII. Tang Y. et al. (2021) Lesion Segmentation and RECIST Diameter Prediction via Click-Driven Attention and Dual-Path Connection. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_32

XIV. Wang J., Xia B. (2021) Bounding Box Tightness Prior for Weakly Supervised Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12902. Springer, Cham. https://doi.org/10.1007/978-3-030-87196-3_49

VOLUME III: Machine learning part 2

I. Daza L., Pérez J.C., Arbeláez P. (2021) Towards Robust General Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_1

II. Chen J., Asma E., Chan C. (2021) Targeted Gradient Descent: A Novel Method for Convolutional Neural Networks Fine-Tuning and Online-Learning. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_3

III. You D., Liu F., Ge S., Xie X., Zhang J., Wu X. (2021) AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_7

IV. van Tulder G., Tong Y., Marchiori E. (2021) Multi-view Analysis of Unregistered Medical Images Using Cross-View Transformers. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_10

V. Thermos S., Liu X., O'Neil A., Tsaftaris S.A. (2021) Controllable Cardiac Synthesis via Disentangled Anatomy Arithmetic. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_15

VI. Xie Y., Zhang J., Shen C., Xia Y. (2021) CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_16

VII. Wang R., Chaudhari P., Davatzikos C. (2021) Harmonization with Flow-Based Causal Inference. In: de Bruijne M. et al. (eds) Medical Image Computing and

Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_17

VIII. Mathew S., Nadeem S., Kaufman A. (2021) FoldIt: Haustral Folds Detection and Segmentation in Colonoscopy Videos. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_21

IX. Gu R., Zhang J., Huang R., Lei W., Wang G., Zhang S. (2021) Domain Composition and Attention for Unseen-Domain Generalizable Medical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_23

X. Liu J., Guo X., Yuan Y. (2021) Prototypical Interaction Graph for Unsupervised Domain Adaptation in Surgical Instrument Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_26

XI. Liu M. et al. (2021) Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_30

XII. Liu Q., Yang H., Dou Q., Heng PA. (2021) Federated Semi-supervised Medical Image Classification via Inter-client Relation Matching. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_31

XIII. Dong N., Voiculescu I. (2021) Federated Contrastive Learning for Decentralized Unlabeled Medical Images. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_36

XIV. Puyol-Antón E. et al. (2021) Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_39

XV. Zapaishchykova A., Dreizin D., Li Z., Wu J.Y., Faghihroohi S., Unberath M. (2021) An Interpretable Approach to Automated Severity Scoring in Pelvic Trauma. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_40

XVI. Marrakchi Y., Makansi O., Brox T. (2021) Fighting Class Imbalance with Contrastive Learning. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_44

XVII. Khakzar A. et al. (2021) Towards Semantic Interpretation of Thoracic Disease and COVID-19 Diagnosis Models. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_47

XVIII. Henn T. et al. (2021) A Principled Approach to Failure Analysis and Model Repairment: Demonstration in Medical Imaging. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_48

XIX. Graziani M., Palatnik de Sousa I., Vellasco M.M.B.R., Costa da Silva E., Müller H., Andrearczyk V. (2021) Sharpening Local Interpretable Model-Agnostic Explanations for Histopathology: Improved Understandability and Reliability. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_51

XX. Larrazabal A.J., Martínez C., Dolz J., Ferrante E. (2021) Orthogonal Ensemble Networks for Biomedical Image Segmentation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_56

XXI. Najdenkoska I., Zhen X., Worring M., Shao L. (2021) Variational Topic Inference for Chest X-Ray Report Generation. In: de Bruijne M. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_59

The impact of GDPR: the researcher's perspective

Lois Riobó, Daniela Varela, Aina Albajar

Master in Computational Biomedical Engineering
lois.riobo01@estudiant.upf.edu, daniela.varela01@estudiant.upf.edu,
aina.albajar01@estudiant.upf.edu

Abstract. In May 2018 the European Union's new data privacy and security law, called the General Data Protection Regulation (GDPR), was put into effect. This new regulatory framework is changing the way in which researchers in the European Union have to deal with personal information in their daily work and can potentially affect the quality of their research outputs. In this study we analysed how researchers consider that the new GDPR is affecting their work in terms of daily practices and productivity. To answer this question, a literature review and a survey were conducted. The survey was distributed among senior researchers in the European Union that work with personal data in different fields. The answers show that, despite researchers recognizing that they have had to make an effort to adapt to the GDPR, they consider that both research quality and rate of production will not be significantly affected.

Keywords: GDPR, personal data, data privacy, scientific research.

1 Introduction

Research is of elemental importance to the prosperity and advancement of any society [1]. It contributes to scientific knowledge, economic growth and can also be used to address societal problems. Interestingly, research largely relies on the use of data, and this often includes personal data. Personal data is data that relates to living people from which they can be directly or indirectly identified. Consequently, access to personal data is often a key factor in determining whether a research project is able to proceed [2]. That is why regulatory frameworks and, specifically, data protection frameworks play an important role in determining what research projects may be conducted [3].

For all the Member States of the European Union the current data protection framework is the European Union's General Data Protection Regulation (GDPR). This new framework is applicable as of May 25th 2018 and regulates the protection of natural persons with regard to the processing of personal data and the free movement of such data. Its aim is to homogenise the rules for all the Member States and to reinforce data subject's rights in a digitalised environment. However, it considers scientific research a specific context of personal data processing where an equilibrium between individual freedom and the freedom of research is required. Therefore, it sets

rules that allow personal data processing and sharing when there is a public interest [4].

Briefly, some of the most important rules the GDPR introduces are the following [5]:

1. Data protection principles: data controllers must ensure that their processing operations are secure, transparent and that privacy is taken into account at all stages of processing.
2. Data subject rights: data controllers must facilitate the right of "erasure" (commonly known as "the right to be forgotten") and the right "to object to processing" among others.
3. Administrative requirements: in most cases, data controllers must appoint a Data Protection Officer and perform a data protection impact assessment.
4. Legal base: data controllers must ensure that there is a legal base for processing, that is, a context in which the processing of personal data is permitted.

One of the questions that arises after the application of the GDPR rules is how they will impact the research field. We believe that getting an idea on how scientific researchers feel about or have been affected by the GDPR will help to identify the real impact of this regulation in their daily work, and such considerations should be taken into account to make further adaptations of this or other laws regarding data protection. Moreover, it may help to evaluate how effective has the divulgation been in this industry and which will be the long-term impact on research production.

Although there are quite a lot of studies regarding the impact of the GDPR on business [6, 7, 8], there are very few previous studies regarding the individual perception towards GDPR. Delloit and Eurobarometer are an example of systematic surveys aiming to explore the impact of GDPR on the relationship between organisations and its clients, and the general awareness among Europeans of the GDPR, respectively [9, 10].

Other studies try to answer the following questions: "What is the user's perspective on GDPR? Do they feel empowered indeed, and did the GDPR succeed in strengthening individual rights and conveying a feeling of confidence and control?" They used a questionnaire administered online by a large research institute in the Netherlands. The results suggest the interviewees (N=1288) are aware of the law and know at least some of the individual rights granted to them. However, they showed reactance and doubt in its effectiveness [11]. Another study concerned with how GDPR is perceived explored how perceptions and attitudes have changed after the first year of compliance, from a marketing perspective [12]. To the best of our knowledge, no effort has been made to explore individual perceptions in academic research.

Therefore, the aim of this project is to answer the following two questions: Are senior researchers that work with personal data in the European Union aware of the GDPR? How do they consider that the GDPR is affecting their work in terms of research quality and rate of production? We hypothesized that senior researchers that work with any kind of human data are aware of the GDPR implications and that their rate of production, not the quality of their research, has been temporarily affected by it. To test this hypothesis, we relied on previous studies in the literature and a survey that we distributed among senior researchers in different research fields.

2 Research methodology

Participants and procedure:

The data used in this study was collected between November 1 and November 15, 2021 from an online survey. As the main goal of the study was to analyse the impact of this legislation in researcher's day-to-day work, other methods such as a scientific production analysis were discarded. This survey consisted of 10 different questions divided in 3 sections: general information, GDPR impact on your research productivity and personal opinion about GDPR. For 6 of these questions a short answer was required and for the other 4, the answer was a quantitative measurement on a 1 to 5 scale. The survey was designed to take about 3 minutes to answer it. It was sent to both teachers and students from different universities, taking into account that it was meant to be answered by people who are currently working with personal data in a research group.

A total of 25 answers were collected during this period of time. Because of the motivation of this work, two conditions were set in order to select the most relevant answers: they had to come from a senior researcher, as junior researchers have not worked with previous legislations in this field, and these participants needed to be working with personal data in the European Union. Out of the 25 participants, only 10 were selected for the analysis due to these restrictions. All the results obtained were anonymized and only general questions about their research work were asked. Distributions in terms of gender, age and education were not taken into account because of the limited number of answers.

Measures:

General personal information was assessed with three questions: 'What is your research area?', 'Which type of personal data do you handle in your research group?' and 'How long have you been working with this type of personal data?'. The main purpose of this section was to classify the answers according to the research field and to discard those answers that came from junior researchers (less than 2 years of experience) or from researchers that do not work with personal data procedures that may be affected by the GDPR.

"GDPR impact on the research work" section was aimed to reflect the degree to which researchers are affected by this new legislation on their daily work in terms of working methods adaptation, efficiency and rate of production. For this reason, four questions were asked, with a 1-to-5 scale answer: 'How familiar are you with GDPR and its impact in scientific research?', 'How would you score the effort that has been made in your research group and/or in your personal work to adapt to the new regulations?, 'Do you think GDPR is affecting the speed at which research outputs are being published?' and 'How is GDPR limiting the ability to obtain the maximum benefit out of your data?'. These last two questions aimed to evaluate the rate of production and the quality of research, respectively. One last multiple choice question was asked: 'If GDPR is affecting your research production, do you think it will be a transient or a permanent effect'?.

Finally, "Personal opinion about GDPR" section was aimed to both obtain a general perception about the motivation of this law ('How well do you think GDPR preserves the equilibrium between protecting personal data while allowing its processing for scientific research purposes?') and to identify specific weaknesses in its implementation ('Is there any particular aspect in which you think this law is ineffective?'). The answers consisted of a multiple choice and a short answer respectively.

Advantages and drawbacks of the model:

The main strength of this approach is that, with an anonymous survey format we can collect a quick perspective of the impact of GDPR in different fields coming from several personal opinions. The main limitation of the project consisted of the limited number of responses that were collected due to both the short period of time during which it was carried out and the relatively low rate of response.

3 Results and discussion

This study describes the results of a panel survey (N=10). The answers for the questions under the General personal information section of the survey were mainly used for the purpose of validating the inclusion conditions for further analysis: researcher with 2 or more years doing research and working with personal data.

The responses for the questions about GDPR impact on the research work are presented in four histograms and one pie chart, Figure 1 to 5, below. None of the participants that met the criteria were completely or very unfamiliar with GDPR (Figure 1), and in general, they scored the adaptations being implemented in their research group with 3 and 4 (Figure 2), meaning that they have been affected by the regulation to some extent, as we have hypothesized. The responses are equilibrated at medium levels on the level of affectation perceived towards the speed at which

research outputs are being published (Figure 3), only one person thinks it is strongly slowing down the production, while the rest equally distribute in levels 2, 3, 4.

The other aspect in which we thought the regulation could be affecting the research was by limiting the ability to obtain the maximum benefit out of the data, however, researchers didn't agree on this matter, 40% opted for "no effect" and only one of them qualified it as "very restrictive" (Figure 4). In the pie chart (Figure 5), we can see the majority (60%) agreed that the impact of GDPR on their research production will be transitive, that is, temporarily and a matter of adaptation. None of them think that the regulation permanently restricts their work.

Figure 1. Bar graph for "How familiar are you with GDPR and its impact in scientific research?"

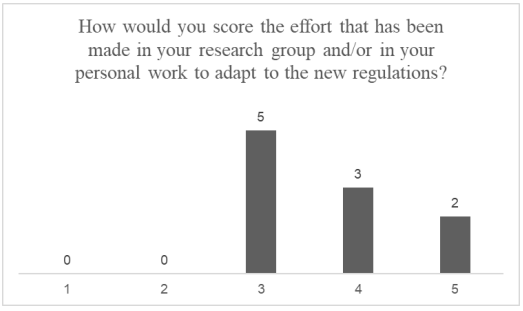


Figure 2. Bar graph for "How would you score the effort that has been made in your research group and/or in your personal work to adapt to the new regulations?"

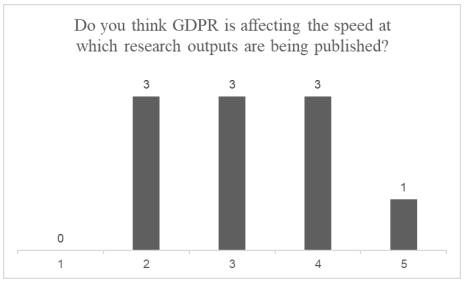


Figure 3. Bar graph for "Do you think GDPR is affecting the speed at which research outputs are being published?"

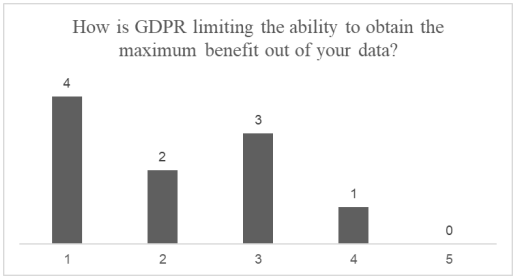


Figure 4. Bar graph for "How is GDPR limiting the ability to obtain the maximum benefit out of your data?

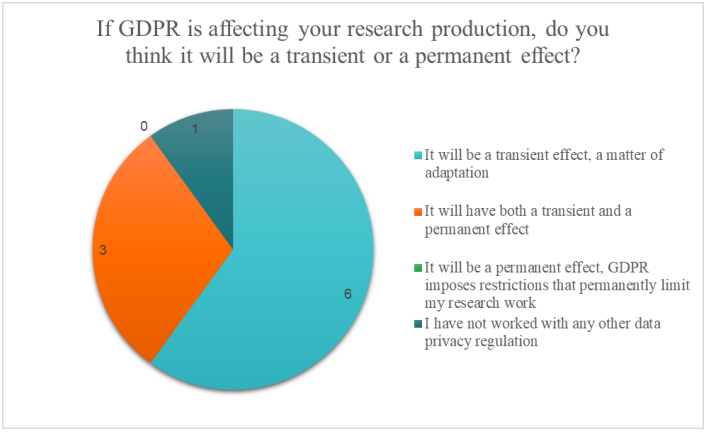


Figure 5. Chart for "If GDPR is affecting your research production, do you think it will be a transient or a permanent effect?"

In the Personal opinion about GDPR section, to the question: "In your personal opinion, how well do you think GDPR preserves the equilibrium between protecting

personal data while allowing its processing for scientific research purposes?" 30% (3 out of 10) responded that it is well balanced, while the remaining 70% agreed that it is more focused on protecting data privacy than on preserving research production. Finally, for the last question: Is there any particular aspect in which you think that this law is ineffective or poorly formulated? not many responses were collected as it was not mandatory, hence we selected only two for illustration purposes:

- "National and regional heterogeneities in data management based on different levels of interpretation of the GDPR, and lack of integrated resources and infrastructures to efficiently share data for research in full respect of GDPR. These aspects are still limiting."
- "It could be unfective as long as researchers do not get to know the overall principles/aims as well as the institutional application for its fulfilment."

Overall, the results indicate that divulgation of the GDPR among researchers has been effective and, moreover, that the GDPR is not perceived as an impediment for research. These conclusions are supported by the results presented in Figure 1, which show that senior researchers that work with personal data in the European Union are aware of the GDPR. In addition, Figure 2 and Figure 3 indicate that adaptation to the GDPR has required an extra effort and has delayed some projects. However, Figure 4 and Figure 5 show that researchers perceive that both research quality and rate of production will not be affected in the long-term.

Therefore, we validated the hypothesis: senior researchers that work with any kind of human data are aware of the GDPR implications and that their rate of production, not the quality of their research, has been temporarily affected by it. However, it is not clear to what extent and which factors are being directly affected. This aspect, therefore, remains to be investigated in future studies. Furthermore, future studies could also include more researchers in the survey, which would strengthen the evidence of the results.

4 Conclusions

From the literature study and the survey conducted, we concluded that senior researchers that work with personal data in the European Union are aware of the GDPR. In addition, although adaptation to the GDPR has required an extra effort and has delayed some projects, researchers perceive that both research quality and rate of production will not be affected in the long-term. Therefore, divulgation of the GDPR among researchers has been effective and, moreover, the GDPR is not perceived as an impediment for research. Overall, we consider that these findings are relevant since there are very few studies in the literature that evaluate the impact of the GDPR from the researcher's point of view. Furthermore, the evaluation of data protection frameworks is of great importance since research is a fundamental part of any society

and it often relies on personal data. Future studies could include more researchers in the survey to strengthen the evidence of the results.

References

1. Mirowski, P., & Sent, E.-M. (2002). Science bought and sold : essays in the economics of science. 573.
2. Heffetz, O., & Ligett, K. (2014). Privacy and Data-Based Research. Journal of Economic Perspectives, 28(2), 75–98. https://doi.org/10.1257/JEP.28.2.75
3. The potential impact of the EU general data protection regulation on pharmacogenomics research. (n.d.). Retrieved November 17, 2021, from https://www.researchgate.net/publication/321668020_The_potential_impact_of_the_EU_general_data_protection_regulation_on_pharmacogenomics_research
4. Chassang, G. (2017). The impact of the EU general data protection regulation on scientific research. Ecancermedicalscience, 11. https://doi.org/10.3332/ECANCER.2017.709
5. Quinn, P. (2021). Research under the GDPR – a level playing field for public and private sector research? Life Sciences, Society and Policy, 17(1), 1–33. https://doi.org/10.1186/s40504-021-00111-z
6. Tankard, C. (2016). What the GDPR means for businesses. Network Security, 2016(6), 5–8. https://doi.org/10.1016/S1353-4858(16)30056-3
7. Garber, J. (2018). GDPR – compliance nightmare or business opportunity? Computer Fraud & Security, 2018(6), 14–15. https://doi.org/10.1016/S1361-3723(18)30055-1
8. Lindgren, P. (2016). GDPR Regulation Impact on Different Business Models and Businesses. Journal of Multi Business Model Innovation and Technology, 4(3), 241–254. https://doi.org/10.13052/JMBMIT2245-456X.434
9. A new era for privacy GDPR six months on Contents. (n.d.).
10. Special Eurobarometer survey on data protection - News - National Data Protection Commission - Luxembourg. (n.d.). Retrieved November 17, 2021, from https://cnpd.public.lu/en/actualites/international/2019/06/eurobarometer-2019.html
11. Strycharz, J., Ausloos, J., & Helberger, N. (2020). Data Protection or Data Frustration? Individual perceptions and attitudes towards the GDPR. European Data Protection Law Review, 6(3), 407–421
12. Tarcza, T.-M., Nemteanu, S. M., Popa, A.-L., & Tarca, N. N. (2019). A Comparative Analysis on Perceptions and Attitudes towards GDPR from a Marketing Perspective. Vision 2025: Education Excellence and Management of Innovations Through Sustainable Economic Competitive Advantage, October 2020, 6793–6803.

Equity in academic publishing: the impact of socioeconomic background on Open Science within Europe

E. Encinas, M. Nakagawa, C. Zatse

Computational Biomedical Engineering Master
eva.encinas01@estudiant.upf.edu, mariana.nakagawa01@estudiant.upf.edu,
christina.zatse01@estudiant.upf.edu

Abstract. Nowadays, everything is related to the economic power of a country. But what is the relevance with Open Science? The aim of this research is to find out whether there is an impact of the socio-economic background on open science, and specifically to the number of publications that are produced every year. In order to find out which is the relevance, we performed research dedicated to two countries in a completely different rank in the GDP (Gross Domestic Product) for the last three decades and compared our results graphically. We came to the conclusion that indeed the impact of the socio-economy is huge, but this is not the only factor that influences the publications' productions.

Keywords: socio-economic background, GDP, publications' production

1. Introduction

It is undeniable that the evolution of science is uneven depending on the geographical location in which we are located. It is not only a matter of advancement in a particular subject, but also a matter of multiple factors, be they economic, social, or even political. We are no longer talking only about differences in the amount of research, the above factors can affect the type and categories of studies that are carried out.

Derived from traditional publishing, new terms have been coined, of particular interest being the so-called '*colonization of information*' [1].

During the last decades, the emergence of the Internet has made it increasingly possible to bring the latest scientific discoveries closer to any part of the world, thus giving rise to a *globalization* or *democratization* of information [1].

1.1 Open Science, Open Data, Open Access

Open science constitutes a new concept in the scientific process, relying on collaborative work and innovative ways of spreading knowledge with the use of digital technologies and new collaborative tools. [2] The elements which shape open science are: open methodology, open source, open data (OD), open access (OA), open peer review, and open educational resources. Particularly for the library and information field, the emphasis is most commonly placed on: Open research data and open access to scientific publications. [3]

The definition of open access is: *the practice of having the ability to have access through the internet to scientific information without being charged and to also to have the ability to reuse it*. By providing the research results and making them available to anyone, the science community, public, and profit industries can profit greatly. [4]

When referring to open data, we mean data that can be published and accessed without being charged or having authorization obstacles. However, nowadays, even if the scientists consider the data that have been published as part of the scientific community, a lot of publishers stick to the idea of copyrights over data and demand permission in order to reuse the data. [5]

1.2 The impact of socio-economic backgrounds on open science.

The aim of this work is to study the impact that geographic location can have on the number and characteristics of open access scientific publications. In addition, it is studied whether this globalization of information is reflected in publication trends.

For this purpose, two European countries with different levels of wealth have been selected to get a first idea of how these factors may have affected the number of publications in the last few years.

Specifically, the study will take into account publications from the last three decades for greater temporal relevance, i.e., publications from 1990 to the current year, 2021.

As mentioned, wealth has been taken into account for the selection of two European countries, based on Gross Domestic Product (GDP).

In addition, we wanted to study the progress in the discipline of astrophysics, due to the fact that we are looking for the study of a subject not related to the professions of first necessity, since it is more common to find publications of this nature.

2. Research methodology

In order to answer our hypothesis, we imported data from Scopus, a multidisciplinary bibliographical database. This database provides a lot of information about publications such as articles, papers, references etc., as well as useful information that can help us assess and measure scientific output. Consequently, it appeared to be a simple and clear tool for us to:

- Create a new database with the target information from the filtered Scopus platform.
- Make quantitative comparisons between the number of publications in the countries of interest.
- Visualization of results, to make comparisons in progression trends and draw conclusions.

2.1 Target selection.

In this work, before we turned to Scopus, we took into account the wealth of each country in Europe based on nominal GDP. After studying their ranking, we decided to choose one country with a high nominal GDP and one who was lower in the rank. As a result, we selected Germany, the country with the highest nominal GDP in Europe and the second highest in the world. On the other hand, the Czech Republic has been chosen as it is usually ranked around 20th place in this index, depending on the year. Despite this difference in the ranking, it should be noted that both countries are considered high income countries, although in the case of the Czech Republic this denomination is quite recent, according to the World Bank (Fig. 1).

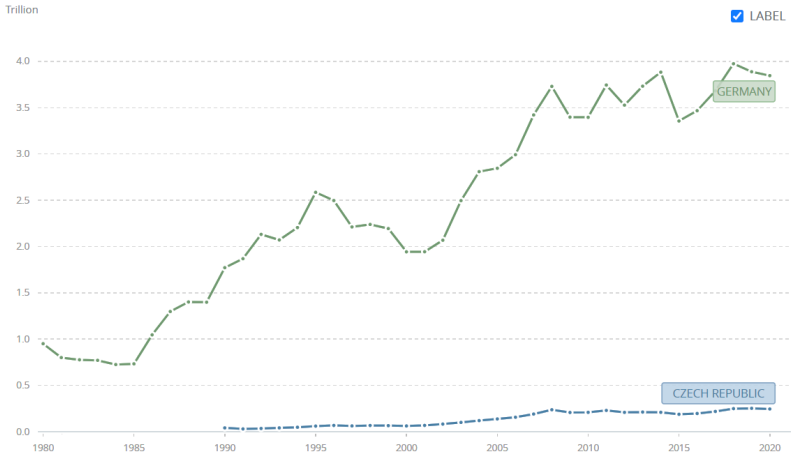


Figure 1. GDP (current US$) - Germany, Czech Republic

We specified our research in the area of Physics and specifically in the field of Astrophysics so as to limit our boundaries as Physics has a lot of branches.

2.2 Filtering information through the Scopus platform.

Having the previous structure in our mind, we executed it in Scopus using the following words in the search:

("Astrophysics") AND (LIMIT-TO(AFFILCOUNTRY,"Germany") OR LIMIT-TO (AFFILCOUNTRY,"CzechRepublic")).

This implies that the filter being applied will return the results of publications associated with the topic 'Astrophysics', in which there is an author affiliated with the countries 'Germany' or 'Czech Republic'.

By conducting the search, a number of publications appeared. In particular 85.873 of publications resulted for Germany, and 8.110 for Czech Republic. Following this, we decided to take the last 3 decades, from 1990 till 2021, into consideration as a timeline in order to have a wide range during the passage of time which could be then analyzed graphically by the tools of Scopus. We ended up with 84.157 publications for Germany and 8.054 for Czech Republic. By taking these data, not only was it possible for us to compare the number of publications between these two countries and define the importance of the wealth factor (according to GDP) to them, but also study the variance of publications during the passage of time and indicate some factors (like big statements) that have led to it.

3. Results

In accordance with the previous process followed, we got the results that are being presented next. In order to better compare trends, data visualization has been used to make progressions over time easier to analyze.

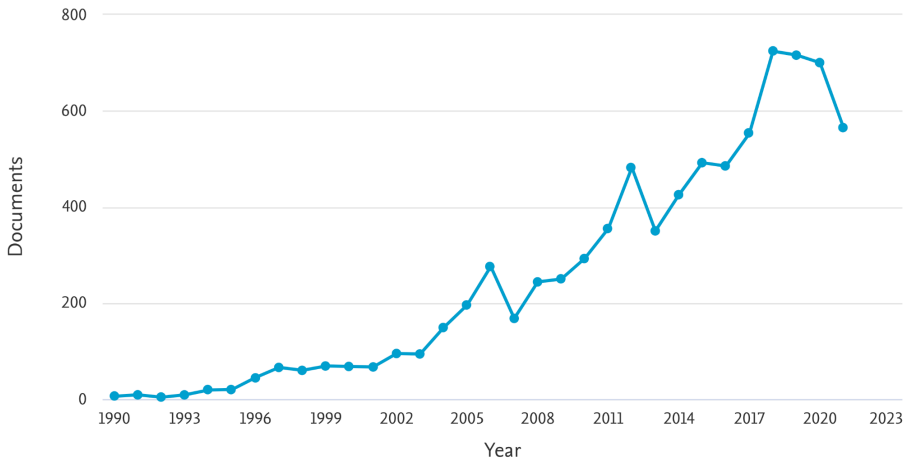


Figure 2. Trend in the number of publications on astrophysics in Czech Republic, during the last three decades.

Figures 2 and 3 represent the number of publications in the Czech Republic and Germany, respectively. The data obtained correspond to the last three decades and it can be seen that, in both cases, the trend is increasing. Despite having similar trends, the beginning of this growth occurs at different times, since in Germany the number of publications begins to increase significantly from 1996 onwards, while in the Czech Republic this happens in 2004, almost a decade later.

The graphs should be contextualized in the political circumstances of the countries prior to these dates:

- In Germany, 1995 is the year in which many authors mark the end of the country's economic reconstruction after the three decades of dictatorial rule (1933-1989) and the withdrawal of Soviet forces from East Germany in 1994.

- In 1993 the dissolution of Czechoslovakia took place and the Czech Republic was formed, after the departure of the communist party from the government in 1989 in the so-called Velvet Revolution. During these years, measures for the economic incorporation into the European framework were implemented, but it was not until 2004 that the Czech Republic entered the European Union.

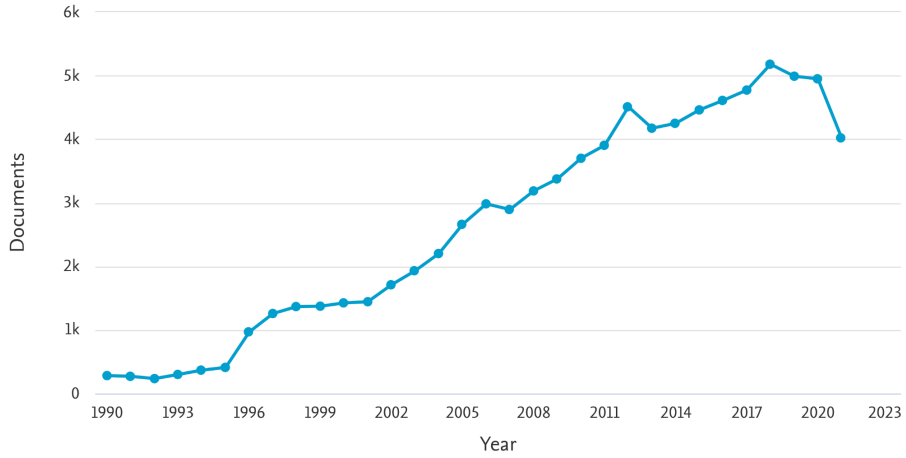


Figure 3. Trend in the number of publications on astrophysics in Germany, during the last three decades.

It is of great relevance to mention the range of values in which the number of publications in each of the two countries moves. While in the Czech Republic the average number of publications per year is at 260, in Germany the average number of publications increases to 2715.

Furthermore, it is worth noting that in both countries, although it is even more noticeable in the Czech Republic (Fig. 2), there are three years in which the number of publications shoots up, standing out from the rest. This happened in 2006, 2012 and 2018, which is due to three major milestones in the history of astrophysics in the last century.

In 2006, Pluto came to be recognized as a minor-planet at the General Assembly of the International Astronomical Union (IAU) held in Prague, Czech Republic, that year [6].

Later, in 2012, graphical evidence for the existence of black holes was published for the first time, records of images of a supermassive black hole 2.7 million light-years away swallowing a red giant were achieved. Also that same year, the deepest optical view of space to date was obtained; in addition to obtaining the most detailed image of the early Universe also known as the first existing light.

Finally, on November 5th 2018 the Voyager 2, travelling in a different direction from Voyager 1, left the solar system, becoming the first interstellar probe [7].The Voyager interstellar mission has the capability of collecting valuable interplanetary, and eventually interstellar, scientific data on fields, particles and waves until about 2025. In 2025, the spacecraft's ability to generate sufficient electricity to keep the science instruments running will expire. [8]

Lastly, both graphs (Fig. 2 and 3) show a clear decrease in the number of publications in the last year. We have assumed that this is due to the health emergency situation in which we find ourselves since the end of 2019, which may have slowed down or even paralyzed the research that was being carried out.

Conclusions

Taking everything into consideration, we can say that the socio-economic background plays a massive role in the publication production. However, that is not the only factor that determines this scientific spread, especially in the field of Astrophysics. We saw that various phenomena, like a pandemic can easily reduce the export of publications (as was noticed a decrease in both countries of our interest, when Covid-19 first appeared). Moreover, we should take into account that we are comparing two countries with a different size and different amount of population (83.24 million people for Germany and 10.7 millions for Czech Republic), which gives a logical meaning to that difference that occured.

In conclusion, when analyzing trends, it must always be taken into account that there are many factors that can affect the facts we see in the data, there is no analysis without contextualization. The situation in a country will undoubtedly affect the research being done in that country, but thanks to globalization there are likely to be trends in common. As seen, big scientific announcements can bring the top down and reach to the peak the number of publications. It seems that, when it comes to science and big “revolutions” , there are no discriminations and all scientists get triggered and try to contribute to any possible extent.

References

[1] Inefuku, Harrison W., "Globalization, Open Access, and the Democratization of Knowledge" (2017). Digital Scholarship and Initiatives Publications. 6. https://lib.dr.iastate.edu/digirep_pubs/6

[2] What is Open Science? Introduction | FOSTER. (n.d.). Retrieved November 16, 2021, from https://www.fosteropenscience.eu/content/what-open-science-introduction

[3] Burgelman, J.-C., Pascu, C., Szkuta, K., von Schomberg, R., Karalopoulos, A., Repanas, K., & Schouppe, M. (2019). Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. Frontiers in Big Data, 0, 43. https://doi.org/10.3389/FDATA.2019.00043

[4] Open access | European Commission. (n.d.). Retrieved November 18, 2021, from https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/open-access_en

[5] Murray-Rust, P. (2008). Open data in science. Nature Precedings, 1-1

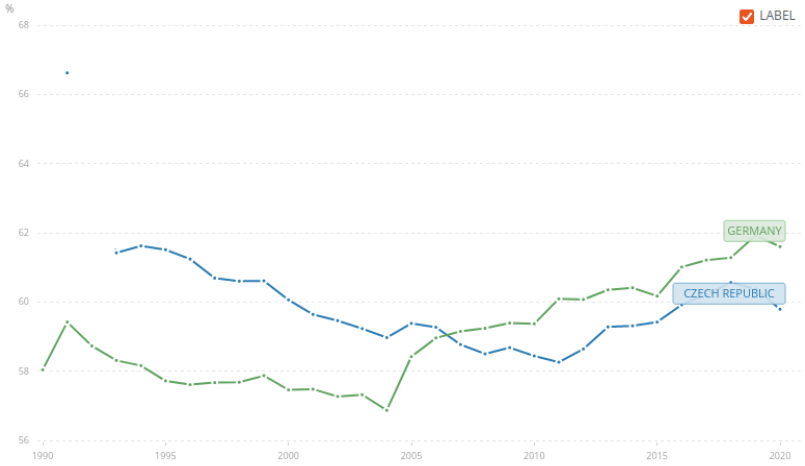
[6] 24 August 2006, International Astronomical Union, https://www.iau.org/news/pressreleases/detail/iau0603/, 15 November 2021

[7] Stone, E.C., Cummings, A.C., Heikkila, B.C. et al. Cosmic ray measurements from Voyager 2 as it crossed into interstellar space. Nat Astron 3, 1013–1018 (2019)

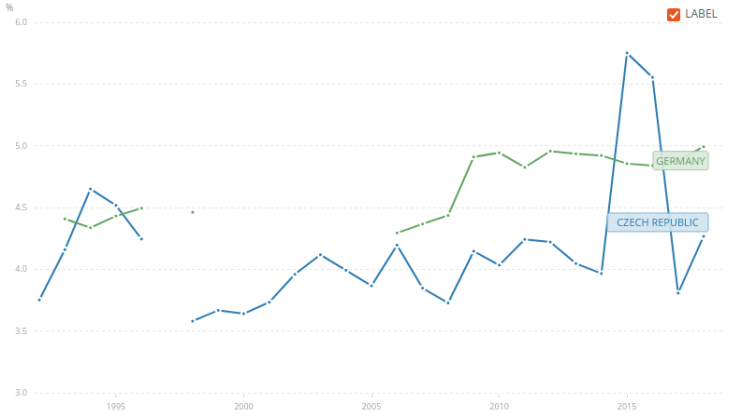
[8] Voyager - The Interstellar Mission. (n.d.). Retrieved November 18, 2021, from https://voyager.jpl.nasa.gov/mission/interstellar-mission/

Appendix

This appendix includes other social and economic metrics with which we have tried to explain the trends obtained in our study but which we have found to have no apparent influence on them.



1.1 Labour force participation rate, total (%of total population ages 15+)



1.2 Government expenditure on education (% on GDP)

How Copyright Restrictions Affect Music Information Retrieval Research

Anna Eliane Barletta, Charles Dean Cochran and Betty Cortiñas-Lorenzo
Sound and Music Computing Master - Univesitat Pompeu Fabra (Barcelona, Spain)
annaeliane.barletta01@estudiant.upf.edu, charles.cochran01@estudiant.upf.edu, betty.cortinas01@estudiant.upf.edu
November 2021

Abstract: In this meta-research paper, we present a modern study on MIR research by reviewing the past of copyright restrictions and identifying the main problems that MIR face as a result of these regulations. We outline the different approaches followed by MIR researchers to overcome these difficulties and the solutions proposed by the community so far to improve the situation in the following years. Then, we present a quantitative analysis to provide evidence on the accessibility and validity limitations that exist in MIR research by analyzing the datasets collected by ISMIR and focusing on the evolution of the number of datasets published from 2000 and the number of full songs. Additionally, we propose a categorization of the types of audio descriptors present in the datasets containing only features. Our results confirm that in the last years, the number of full songs openly available for the MIR community has increased, however, validity of MIR research is still hindered by the huge difference in the number of full songs between open datasets and private ones. We also confirm by statistical analysis that the amount of datasets has noticeably increased, especially the feature-based ones. Finally, we propose a categorization of feature-based datasets and conclude that the types of audio descriptors present are very varied in nature and also reflect the latest advancements in MIR algorithms regarding musically-relevant data extraction.

Keywords: Music Information Retrieval, Copyright, Restrictions, Limitations, Accessibility, Reproducibility, Validity, ISMIR, Datasets, Information Sources, Pubic Domain, Creative Commons

1 Introduction

Music Information Retrieval (MIR) has been defined as "*the research field which focuses on the processing of digital data related to music, including the gathering and the organisation of machine-readable musical data, development of data representations, and methodologies to process and understand such data*" [1]. In other words, MIR tasks seek to extract information (using computational methods) that is contained in music, usually in the form of an audio recording or notation [2]. Examples of typical MIR tasks include: extracting musical features and properties like genre, estimating music metadata, manipulating musical sequences, or synthesizing new melodies following a certain compositional style [2]. Since its beginnings, the MIR field has provided many impactful reports, services, and new observations to the

music industry, which allowed it to develop new technologies for organising, discovering, retrieving, delivering, and tracking information related to music, as well as services for digital media stakeholders. Research within the MIR community is held to a high standard, especially the research presented at the International Society for Music Information Retrieval (ISMIR) [3]. The close-knit community of MIR researchers and organizers have to make sure that the validity of their research is held to the same standard as their peers. Without the ability to verify each other's work, which is a pinnacle step in the scientific process, consensus on new findings is difficult to reach [4].

This process, however, has been hindered by the use of copyright-protected data since the very conception of MIR research. It is well-known that the most significant source of musically relevant data for MIR tasks comes directly from audio content or any other type of information that is computed from the audio file usually referred to as "features" [1]. At the start of the 21st century, however, there were "*no community-wide music collections against which researchers could cross-evaluate a wide variety of different techniques*" [5]. These data accessibility issues in MIR research due to copyright-protected datasets have been acknowledged by several authors over the last years. This situation has led to a big separation between the MIR industrial research, which has access to copyrighted material, and the more academic or independent MIR research, which has limited funding and utilizes public domain resources. As a result, MIR authors have acknowledged the need for better cooperation between private industry and academia that can lead to more datasets available for the MIR community [6].

Accessibility problems have also an impact on two fundamental qualities that good research must meet: reproducibility and validity. On the one hand, reproducibility in research ensures that the outcomes are reliable. Research results that cannot be reproduced are not valuable at all. Usually, differences in implementation can produce deviations in results. Specifically, in MIR research, these issues mainly appear as a consequence of the accessibility problems regarding the datasets. If researchers use private or copyright-protected music datasets, the reproducibility will be in danger, since these datasets cannot be legally shared between researchers, therefore complicating the proper re-evaluation for reproducibility. In fact, it has been pointed out that when this happens studies rarely present results that can be compared with other research, meaning that some papers report overall results without reference to any common measure of significance [5]. Moreover, this privatized and copyrighted material represents most of the existent music datasets for MIR. On the other hand, copyright regulations also affect validity of MIR research, and thus its generalizability. Copyright laws lead to limited access of MIR researchers to musical material. As a consequence, restricted and biased subsets are created which are very "*difficult to generalize to larger music populations"*. In the MIR field, academia researchers over time have been forced to use publicly accessible music datasets, which are often limited in size and skewed towards Western music, especially classical music [4]. As a consequence, MIR organizations like ISMIR have acknowledged over the years the importance of discovering a solution for creating the functionality for researchers to utilize data that can increase the reproducibility and validity of research [2]. In response to this, the MIR community has presented a mass

of software tools, dataset collections, and other useful academic resources for public use.

We understand that the discussion of the limitations of copyright in MIR research is a concept all MIR researchers should be aware of, however, not many papers cover this specific topic in detail, especially in quantitative ways. Therefore, our purpose is to develop in this study a constructive analysis of the effect that copyright restrictions have placed on the MIR research community through an analysis of the specific limitations of accessibility and validity in MIR research. In particular, we aim at analyzing the change in number of full songs available in open MIR datasets over the last two decades and their comparison to private datasets with the purpose of identifying how the validity problem has evolved in terms of the size limitation of datasets. Furthermore, in order to quantify the accessibility problem, we aim at analyzing the evolution in information source contributions and publications since the establishment of the most popular MIR conference ISMIR, distinguishing between audio-based and feature-based datasets. Finally, given the noticeable popularity of feature-based datasets in the MIR research community, we will propose a categorization framework of the features present in those datasets in order to explore the main trends and the current state of the overall feature-based corpora.

We will first provide in Section 2 domain information to understand how copyright restrictions have affected MIR research in the past then and in Section 3 we will explain the research methodology carried out. In Section 4, we will present the results of the analysis on modern MIR datasets to analyze the change in accessibility and validity in MIR research as a consequential effect of copyright restrictions. Finally, in Section 5 we present the conclusions.

2 Background

For the sake of simplifying this overview, we focus on the US copyright laws as an illustrative case, but attention should be given to the fact that copyright regulations are country-specific. In addition, in this section we explain the different approaches carried out by MIR researchers to overcome copyright issues and some solutions proposed in the literature to improve the current situation.

US copyright laws date back to 1790 when the Copyright Act was established. As stated by the authors in [8], "*copyright is a property right ascertained to the author of an original work, such as a literary work of a musical work, which deprives others from engaging in certain uses of that work, for a defined period of time, without the author's consent"*. Setting the scenario for the first ISMIR conference in 2000, the Digital Millennium Copyright Act (DMCA) of 1998 established a set of international mandations aimed at preventing unauthorized access and use of creative works on the internet [4]. These past copyright regulations have therefore prevented MIR research from having adequate data collections since its very beginning. More recently, in 2018 the Music Modernization Act (MMA) was passed to establish a system for music licensing of digitally distributed music [7]. In fact, as of January 2021, all major music streaming or digital signal processing platforms are now required to report playback usage to The Mechanical Licensing Collective [9].

On the other hand, over the last twenty one years, society's access to music datasets have tremendously changed. Shift in published research and increase of academic presence in MIR could be attributed to the creation of ISMIR, being the most well-known conference by the MIR community, and established in 2000. ISMIR has been growing in popularity and participation since its first conference, which has encouraged the MIR community to develop a growing collection of audio collections for the purpose of MIR research. Though there still seems to be a separation between industry and academic research, MIR researchers have created several information sources using Creative Commons licensed music recordings; most of which use the *Creative Commons BY-NC-SA* (*Attribution-NonCommercial-ShareAlike)* license. This license is ideal for academic and independent research since it allows music to be shared, copied, redistributed, without it being used for commercial or redistribution purposes. Some of the most notable data collections created in the past twenty years would be the Million Song Dataset (collection of metadata and precomputed audio features for a million songs), MedleyDB, AcousticBrainz, and Free Music Archive, which have been extensively used in MIR research. However, many researchers are using private commercial and in-house datasets owned by companies like *Spotify*, *Deezer*, *Last.fm*, and other "for profit" companies. These usually require a significant financial investment. There are some other modalities of datasets that exist which are not as frequently used, but are still well known in the field of MIR. The modalities come in the form of datasets that have been aggregated from others, like the RWC dataset, built as the world's first large-scale music database for MIR research purposes. Other data collections for research also exist like the MuMu dataset, a multimodal music dataset that combines meta-data from the Amazon Reviews dataset, which contains metadata gathered from Amazon.com, and the Million Song Dataset.

To extend on the limitations specifically created by copyright restrictions, modern music is mainly privatized and copyrighted. Therefore it cannot be legally shared for research purposes. This as previously stated has limited reproducibility and validity capabilities in MIR research. As noted above, MIR researchers have been forced to use publicly accessible Creative Commons and public domain datasets, which the main drawbacks are the bias towards Western classical music of this type of data and the limitations in size. This is a primary cause of variations and weak experimental power in MIR research [4]. Non-Western genres remain comparatively under- studied versus pop, rock and dance music.

The use of publicly accessible Creative Commons and public domain datasets has been one of the main approaches used by MIR researchers to overcome the copyright limitations. Another solution is the use of datasets composed of audio descriptors that do not include the audio files. This has been a great proposal and has been extensively used during the past years, in fact, the authors in [4] state that one fifth of all datasets used in ISMIR 2018 are of this type. However, it is not all benefits, this kind of datasets do not allow for an in-depth analysis since researchers are limited to the provided features and cannot access the audio itself. Other solutions include the proposals of the authors in [4]. They have proposed that developing a distributed MIREX system for making comparisons between MIR algorithms would allow researchers to test algorithms without violating the copyright or relying solely on non-copyrighted music [10]. These authors also proposed a Researcher API

License that could leverage existing industrial infrastructure. In this way, "*companies with existing developer licenses and API infrastructure could create a new license that allows for common MIR practices and explicitly prohibits using algorithm outputs for commercial or listening purposes, but limits casual listening"*.

3 Research Methodology

It has been already established that in modern MIR research, obtaining large multimedia collections for widespread evaluation is less challenging than it has been in past years [9]. This could be attributed to the presence of ISMIR and the creation of new Creative Commons licensed MIR data collections. In response, it should be possible to analyze how copyright restrictions affected the accessibility of MIR research, by analyzing the change in openly accessible MIR information sources, and analyzing the distributions of audio/feature datasets. Verifying whether the number of information sources in recent years has significantly changed would highlight the effect on accessibility of MIR research that copyright restrictions had created. Also, verifying whether the distribution of audio based and feature based information sources in recent years has significantly changed would additionally highlight the effect on accessibility of MIR research. In order to show this, we perform a quantitative analysis on the evolution of published databases of MIR. For the quantitative analysis proposed, we collected our source material from ISMIR website [3], which is a credible, non-exhaustive list of data collections that represent the MIR community's research. The list that was collected spanned from 2000, the year of the first ISMIR, to the present, and included the following features: the name of the dataset, the type of metadata present in the dataset, the specific contents and the presence or not of audio files in the dataset. In addition to the source material, we accessed each of the publications corresponding to each of the provided datasets and included an additional feature representing the publication date of each of the datasets.

Prior to analyse the impact of copyright regulations on the accessibility of data for MIR, we explored the impact on validity by performing a simple analysis, especifically attending to the size limitation of the datasets, which is one of the main causes of validity issues. Because of lack of time, we had to limit this analysis to the number of complete songs available for the decades 2000-2010 and 2010-2020, and could not expand our analysis to cover other factors affecting validity such as the presence of biases in the datasets. The validity analysis was done by extracting the number of full songs of the analysed datasets. Specifically, we explored the audio-based datasets and manually selected the ones containing full songs or compositions, that is, complete audio files of more than 30 seconds long with relevant musical content. As well as discarding the datasets containing snippets, we also did not consider datasets containing audio corresponding to mixes, technical exercises or just notes or simple musical phrases, beats, or patterns. We summed the number of complete songs of the datasets proposed during decade 2000-2010 and decade 2010-2020, and compared these numbers between them and to the number of songs available in common private datasets. This will provide a simple insight on the validity issue regarding the size limitation of datasets.

As for the accessibility analysis, utilizing state-of-the-art Python libraries *pandas*, *matplotlib* and *seaborn* to explore the quantity of information sources for each year from 2000 to 2020, we proposed a statistical test to identify whether the distributions of audio-based and feature-based information sources in recent years has changed when compared to the number of information sources available at the establishment of the ISMIR. It is not likely that one could form a complete list of all MIR datasets that have been utilized in MIR research, due to the sheer size of information sources that are obtainable through the internet. So for the purposes of this meta-research analysis, we considered the collection provided by ISMIR, the most prominent conference in the MIR field, to be a strength to this approach. A weakness of this particular approach is the manual collection of each publication date. In addition to this, the data utilized is only partially outdated since some of the versions of information sources have been updated by their publishers.

The statistical method of our choice to identify whether the distributions of audio-based and feature-based information sources in recent years has changed when compared to the number of information sources available at the establishment of the ISMIR is the Chi-Squared test. For the purposes of this study, we validated the following assumptions in our data to utilize the Chi-Squared test. We also assumed here that our dependent variable, being that number of publicated data sets, is measured at a continuous level. As for our independent variable, we assumed that the audio-based datasets published are independent of those feature-based datasets in the years of 2000 to 2020. Additionally, we assumed that our observations are independent of each other. This means that no single publication is observed in both audio and feature based distributions.

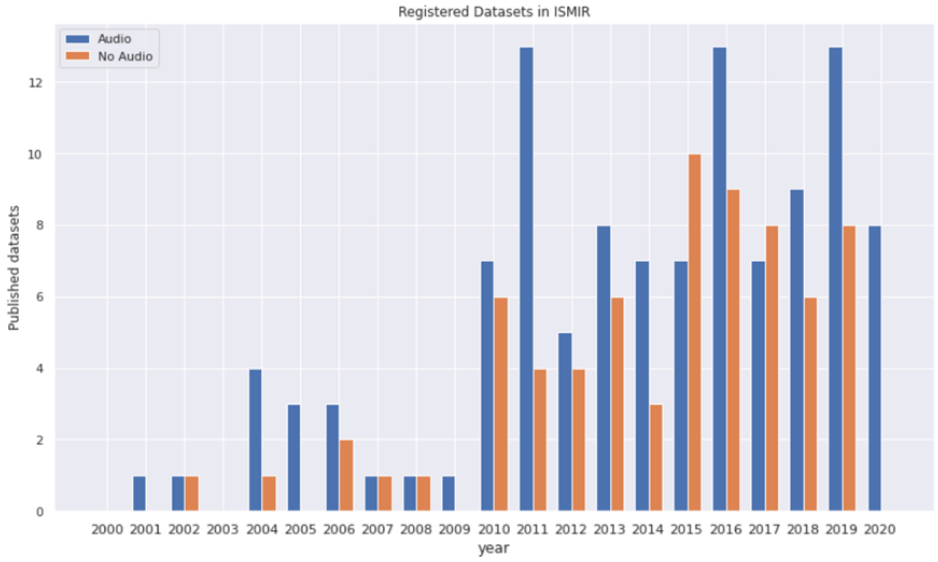


Fig.1 : MIR published datasets according to ISMIR from 2000 to 2020

4 Results

4.1 Validity

Table 1 shows the number of full songs available in the analysed datasets for the two last decades. As we can see, the number of full songs has increased noticeably in the last decade. However, a more detailed analysis shows that this increase comes fundamentally from the creation of two open large-scale datasets in MIR: FMA-full with 106574 full songs and MTG-Jamendo with 55701 full songs. Although this increasing number throws encouraging prospects on the evolution of MIR in terms of validity of its research (and also to the accessibility), the numbers are still very low compared to the number of full songs available in private repositories, like Amazon Music Unlimited dataset with 2 million full songs or Apple Music dataset with 45 million full songs.

2000-2010 decade	2010-2020 decade
459 full songs	180392 full songs

Table 1: number of full songs in MIR open datasets proposed over the two last decades

4.2 Accessibility

From our methodologies we were able to explore and understand the various information sources that were published in the past 20 years. In doing so we were able to visualize the increase of publicated datasets in Figure 1. This led us to believe that there could be significant information for us to conclude upon, seeing that the information present in the decade after 2010. Analyzing the frequencies of these publications helped us identify that there was an approximate 480% increase of feature-based datasets and a 420% increase in audio-based datasets aggregated in the last decade when compared to the years 2000 through 2010. Aside from this, of the datasets published from 2000-2010, roughly 35% of the datasets contained only feature values, while the other 65% contained audio.

In addition to analyzing the statistical increase of music datasets, we utilized a Chi-Square test to verify whether the distribution of audio-based and feature-based information sources in recent years has significantly changed. The null hypothesis was that datasets containing audio for MIR research and the decade they were published in are independent. Through our measurements of the Chi-Squared test. The p-value was numerically represented by 45.2%. This was indicative that we should not reject the null hypothesis at 95% level of confidence.

4.3 Categorization of features

Furthermore, we presented a percentage distribution of the audio descriptors categories present in the descriptors-based datasets that we collected from ISMIR. The results show that metadata is the most common type of descriptor, closely followed by important musical features regarding harmony, structure and rhythm, thus showing the increased progress of MIR algorithms in recent years for extracting musically-related descriptors. We summarised the possible categories, and come to the following classification:

- Emotion, perceptual and expressive features: can include tags about emotional content of music like arousal and valence, perceptual and psycho-acoustic information as well as expressive features, such as the presence of vibrato in the music recording.
- Listening habits: annotations describing behaviours and patterns of users when listening to music.
- Metadata: can include features describing the name of the work/piece/song, artist and composer, timestamps of recording and publishing, era, instrumentation and genre.
- Rhythmic: can include features describing rhythmic content, such as beats, timing, tempo and measures.
- Structural: can include features describing structural or formal characteristics of the music, such as parts of the song/piece, location of cadences, onsets, certain motives and existence of melodic/harmonic/formal patterns.
- Harmony and key: can include features describing harmony aspects of the music, such as chord and key analysis.
- Notes and melodies: can include analysis and identification of notes and certain predominant melodies in the song/piece.
- Lyrics: can include the lyrics of the songs themselves or important information about them.
- Text information: additional textual features like biographical information of the composers or artists, and reviews of the music.
- EEG: features extracted from electroencephalographic studies.
- Others: other types of features included as annotations that do not fit into the above categories. These include: symbolic scores, aligned MIDI and sheet scores, Optical Music Recognition features, similarity with respect to other songs/pieces, popularity ratings, texture analysis and general tags about theme or musicological characteristics.

Fig. 3 shows the percentage distribution of these audio descriptors of the “no audio” ISMIR datasets collected. As it can be observed, “Others” category is the most predominant, meaning that, in fact, annotations in these datasets are quite variant in nature. Even so, we can see that “Metadata” is quite common, as well as “Structural”, “Harmonic” and “Rhythmic” audio descriptors are quite common.

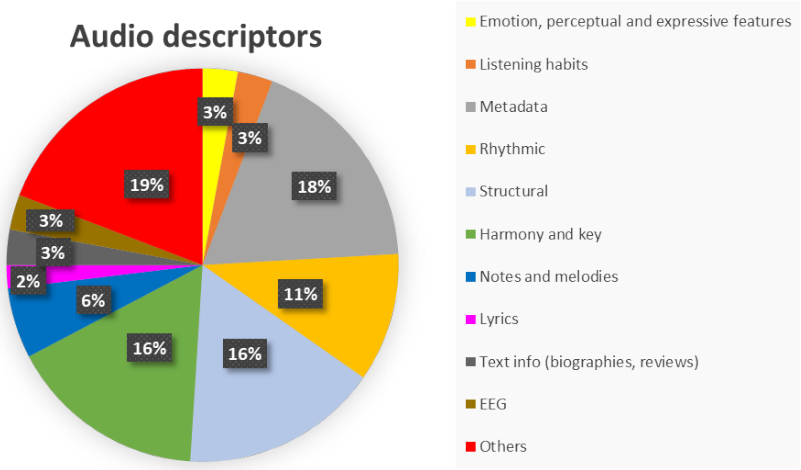


Fig.3 : Audio descriptors categories distribution for ISMIR datasets published from 2000 to 2020

The latter categories suggest that MIR advancements in detecting harmonic, structural and rhythmic features is influencing the number of datasets including this type of information, which is a positive fact. However, we found that only 1 dataset was referring to "Texture" information and also very few were exploring the "Notes and melodies", which we consider is a very important type of information that can be analysed in a musical piece, together with its harmony, rhythm and formal structure. Emotion and perceptual features are also under-represented in the datasets we analysed.

5 Conclusions

The discussion of the restrictions of copyright in MIR research is a concept all MIR researchers are very aware of, however not many papers address this specific topic in detail. In this meta-research paper, we have provided evidence on the existence of those restrictions by referring to the literature on the approaches followed in MIR research to overcome the effects of these copyright laws on data accessibility, reproducibility and validity. We identified the solutions proposed by the MIR community to face a better future for the field.

In order to further provide evidence on the effects of copyright restrictions on accessibility and validity of MIR research, we conducted a number of analyses using the datasets provided by ISMIR website between the years of 2000 and 2020. We extracted the number of full songs available in the analysed audio-based datasets and confirmed that, although new efforts we created in the last decade to create new open datasets with raw audio available for MIR research, the numbers are still far from ideal to be fairly comparable to the available datasets for private MIR research (i.e. based on private or copyrighted material). This simple analysis has shown the

existence of validity issues in MIR research. Because of time limitations, we could not explore other factors affecting the validity in MIR research.

Furthermore, we presented an overview analysis of the evolution in dataset publication since the beginning of ISMIR and confirmed the increase of dataset creation initiatives in the last years, for both audio-based and descriptors-based datasets. As we were to also verify whether the number of information sources in recent years has significantly changed, it should be now confirmed that the accessibility of MIR research has not been hindered by copyright restrictions significantly in the past 10 years. This was observed as a rough 500% increase in publicly accessible datasets over the last decade when compared to the years of 2000 through 2010, regardless of the meta-content provided in each dataset. The results of our dataset analysis have also allowed us to interpret that there has not been a statistically significant correlation of the frequency in the number of audio-based datasets, and the frequency in the number of feature-based datasets in the last two decades. We can conclude that since there is not a significant correlation, that the frequency of the creation of audio-based datasets has not been significantly limited in the years of 2000-2010 when compared to the years of 2010-2020. This suggests that if researchers were to study MIR there would not be a significant difference in the accessibility of audio-based datasets than if they were to have studied MIR between 2000 and 2010.

5.1 Future Work and Challenges

Because of the complexity of this topic, we reiterate the need for all MIR enthusiasts to comprehend this paper. Ignoring these limitations hinders the underlying evolution of MIR study. As for the work that could further analyze the results present in this paper. Further studies on the validity problem could aim to explore in more detail the biases in the open datasets to provide a better understanding of the problem. On the other hand, observing the changes in published information sources in the upcoming decade would be beneficial to the MIR community to further verify whether there is a correlation between the frequency of information sources and the given decade. As for the future challenges that MIR research faces in regards to this paper, research will still be in a deficit of audio collections. However as previously mentioned in this paper, there has been a significant increase in plenty of datasets of different types. Having audio for your MIR research is no longer a requirement and or a concern that many MIR researchers are focused on; for now at least.

References

[1] Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., & Widmer, G. (2013). *Roadmap for Music Information ReSearch*.

[2] Holzapfel, A., Sturm, B. L., & Coeckelbergh, M. (2018). Ethical Dimensions of Music Information Retrieval Technology. Transactions of the International Society for Music Information Retrieval, 1(1), 44–55. DOI: http://doi.org/10.5334/tismir.13

[3] ISMIR. 2021. *ISMIR*. [online] Available at: <https://ismir.net> [Accessed 19 November 2021].

[4] Chen, W., Keast, J., Moody, J., Moriarty, C., Villalobos, F., Winter, V., Zhang, X., Lyu, X., Freeman, E., Wang, J., Cai, S., & Kinnaird, K.M. (2019). Data Usage in MIR: History & Future Recommendations. ISMIR.

[5] Futrelle, Joe, and J. Stephen Downie. "Interdisciplinary communities and research issues in Music Information Retrieval." *ISMIR*. Vol. 2. 2002.

[6] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. 2006. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 2, 1 (February 2006), 1–19. DOI:https://doi.org/10.1145/1126004.1126005

[7] 115th Congress. H.R.5447 - Music Modernization Act. https://www.congress.gov/bill/115th-congress/house-bill/5447/text. [Online; accessed 01-November-2021]

[8] Karydi, Dimitra & Karydis, Ioannis & Deliyannis, Ioannis. (2012). Legal Issues in Using Musical Content from iTunes and YouTube for Music Information Retrieval.

[9] Music Licensing Transformed by the Passage of the Music Modernization Act. SR Englund, AI Stein, AU Mcalpin - Communications Lawyer, 2019 [Online].

[10] J. S. Downie, X. Hu, J. H. Lee, K. Choi, S. J. Cunning-ham, and Y. Hao. Ten years of MIREX: Reflections, challenges and opportunities. Proc. of 15th ISMIR Conference, pages 27–31, 2014.

Impact of Open Access to Datasets: A Case Study of Indian and Chinese Traditional Music

Huicheng ZHANG*, Yuxi QIAO*, and Qingyuan LIU*

Master in Sound and Music Computing, Universitat Pompeu Fabra
{huicheng.zhang01, yuxi.qiao01, qingyuan.liu01}@estudiant.upf.edu

Abstract. The accessibility of music dataset is crucial to the research field of Music Information Retrieval. By applying bibliometric analysis to data retrieved from Google Scholar and Scopus in the period of 2000-2020, we focused on the impact of selected dataset-creating paper within the research field of Indian art music and Chinese traditional music. We found that the publishing time of papers from a open dataset project correlated to the sudden increase of popularity of the corresponding research field. These publications of open accessed dataset and authors tend to have more impact to the field. In a word, high accessibility of music dataset can lead the research field to high prosperity.

Keywords: Meta Research · Open Access · Indian Art Music · Chinese Traditional Music.

1 Introduction

In recent 10 years, machine learning has been developing rapidly. Besides improved algorithms and computational power, the increased availability of large datasets provides a material foundation for machine learning techniques developing[1], [2]. Insufficient dataset is seen as a factor that limits the progress in algorithms[3]. As machine learning has gradually turned into a data-driven field, datasets had been the key role to orient the goals and values of the research field[4].

Various researches have been done on the impact of open access. The majority of them focus on the open access *of papers*, some suggest open

* Joint first authors with equal contribution

accessible papers have an advantage on citations[5]–[9], while there are also some opposite opinions[10]–[12]. However, few researches attempted to study the impact of open accessible datasets. Wenqin Chen et al. conducted a related research, in which they inspected the accessibility of datasets used in Music Information Retrieval field, and stated that "Unequal access to music data has led to field-wide issues including a crisis of reproducibility and concerns about access"[13].

Bibliometric (or scientometric) analysis are favored when doing meta analysis on academic publications[14]. According to E. Abdeljaoued et al., "Bibliometric analysis can be defined as applying statistical techniques on a collection of publications retrieved from a large academic database to analyze and understand the global research output in a particular field" [15]. Bibliometric analysis investigates citation relationship, co-occurrence of keywords, co-authorship and other metric, which can be helpful to measure the impact of some selected publications or research topics. In our case, we focus on the impact of papers that create datasets (dataset-creating papers) within the collection of papers in the related domain (domain papers), based on the bibliometric data retrieved from Google Scholar and Scopus in the period of 2000-2020. Dataset-creating papers were several papers selected manually, which introduced a new dataset to the field. Domain papers were publications collected in the database using specific domain-related query string.

In the last decade, many datasets of Indian and Chinese traditional music emerged, along with a research trend of approaching research challenges from a culture specific perspective, which inspires us to carry out a quantitative research on the impact of those music datasets. CompMusic project[16] is a good example of open science, which provides us several potential music traditions to study, namely Hindustani (North India), Carnatic (South India), Turkish-makam (Turkey), Arab-Andalusian (Maghreb), and Beijing Opera (China). In this paper, we investigated two regional music traditions: Chinese traditional music and Indian art music. We considered them because they are both music traditions that have not been fully explored, unlike heavily favored western classical and popular music. When

selecting the dataset-creating papers, we pay special attention to the publications from CompMusic project. The detailed methodology and results will be discussed in the following sections.

2 Methodology

2.1 Scope of Study

The scope of study depends on the total number of retrieved publications and types of sources. In order to study the impact of the openness of datasets to the corresponding field, some definitions need to be clarified here. We define "dataset-creating papers" as a collection of papers which introduced a newly built dataset. The list of dataset-creating paper can be found in the reference [16]–[18]. In terms of "the corresponding field", we regard it as a large collection of publications which contains specific query strings in title, abstract or keywords. The query strings for collecting domain paper in Scopus are shown in the table 1. We focused on the bibliometric data in the selected databases.

Martín-Martín et.al. suggested that in all areas Google Scholar citation data is essentially a superset of Web of Science (WoS) and Scopus, with substantial extra coverage[19]. However, Google Scholar limited the way we query, and we are forced to use relatively simpler queries, and manually filter the results fetched. The query strings used for collecting domain paper on Google Scholar are shown in table 3

2.2 Data Gathering and Processing

We use various ways to gather the data needed. The major data sources we used are Google Scholar and Scopus.

Google Scholar

We first construct a query to search on Google Scholar, with the help of Publish or Perish, all results are stored locally. Then, we filter out some publications we do not interested in (those without being cited, without publish year record, those books, and some publications we considered out

Table 1. The query strings for retrieving domain paper on Scopus.

Music Tradition	**Query String**
Indian art music	TITLE-ABS-KEY(("chinese traditional music" OR "chinese national music" OR "PIPA" OR "guzheng" OR "guqin" OR "erhu" OR "chinese instrument" OR "jingju" OR "beijing opera") AND ("computational musicology" OR "computational" OR "musicology" OR "deep learning" OR "automatic classification" OR "automatic detection" OR "music information retrieval" OR "data-driven")) AND PUBYEAR > 1999 AND PUBYEAR < 2021 AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ar")
Chinese traditional music	TITLE-ABS-KEY-AUTH(("Indian classical music" OR "Indian art music" OR "tabla" OR "saraga" OR "mridangam" OR "Hindustani" OR "Carnatic" OR "raga") AND ("computational musicology" OR "Automatic classification" OR "automatic detection" OR "music information retrieval" OR "data-driven")) AND PUBYEAR > 1999 AND PUBYEAR < 2021 AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ar")) AND (EXCLUDE (SUBJAREA,"MEDI"))

of topic). We use a crawler to recursively fetch data from Google Scholar according to the reference relationship, starting with a root paper we want to research, then store the reference graph into local SQLite database with the help of python scripts. On that reference graph, we execute various calculation to demonstrate the impact of open-accessible datasets. The pseudocode of the recursively fetching procedure is demonstrated in code block 1.1

Scopus

After searching the query strings in Scopus, 81 publications on Indian art music and 22 on Chinese traditional music were retrieved. By adding dataset paper into the retrieved list, we have 83 publications on Indian art music and 23 on Chinese traditional music, which is what we called the domain

Table 2. Citation information of Dataset-creating paper

Music Tradition	Dataset References	Public	Impact
Indian art music	Serra X. 2014 [16]	Yes	23/29
	Srinicasamurthy et al. 2014 [17]	Yes	10/10
	Chordia et al. 2008 [18]	No	11/16
Chinese traditional music	Serra X. 2014 [16]	Yes	23/29
	Rafael Caro Repetto and Xavier Serra 2014 [20]	Yes	13/15
	Xiaojing Liang et al. 2019 [21]	No	0/1
	Zijin Li et al. 2019 [22]	No	1/2
	Yusong Wu et al. 2019 [23]	No	0/1

Table 3. The query strings for retrieving domain paper on Google Scholar.

Music Tradition	Query String
Indian art music	`"Indian art music" MIR`
Jingju (Beijing Opera)	`jingju MIR`

paper. Comma-Separated Values (CSV) files were downloaded from Scopus for statistics analysis and visualization in python and VOSviewer[24].

VOSviewer

VOSviewer is a software tool for creating maps based on network data and for visualizing and exploring these maps. VOSviewer can parse the CSV file exported from other databases, e.g. Web of Science, Scopus, Dimension and so on. In the graph, the size of circle or frame is related to the size of the user-assigned attribute. The thickness of the connecting line represents the number of co-occurrence times of the end objects. User can modify the scale, change the color and rotate the network to get the satisfied result. VOSviewer has been used widely in the field of bibliometric analysis.

Code block 1.1. Pseudocode of data fetching script

```
job_queue = queue()
job_queue.append(root_article)
while job_queue is not empty:
    job = job_queue.pop()
    articles = fetch_reference(job) # Fetch all articles that cited
        current article
    for article in articles:
        if article is not in domain_publications:
            continue # Ignore those not in our domain publications
                database
        store_ref(job, article) # Store the ref relationship
        if article has never be in job_queue:
            job_queue.append(article)
```

2.3 Metrics

In this paper, analysis will be based on the following metrics: Impact, popularity and co-citation relations and keywords co-occurrence relations. The definition of "impact" is adapted fro Aragón [25], which is defined as the number of citations from other impacting researches. In our case, we consider the publications with more than one citation as impacting researches. We calculated the number and ratio of impacting researches and total number of citations as a measurement of impact (See in the fourth column of Table. 2). It is a good way to compare the quality of citations so that we can have a better understanding of the impact of a paper to the corresponding field.

Popularity is defined as the total number of publications within the research field.

Co-citation is defined as the frequency with which two documents are cited together by other documents[26]. To put it another way, co-citation relation shows how many researches are based on these two publications, which can be considered as a complementary metric of Impact, or "joint impact" more precisely. The larger the number of publications by which two publications are co-cited, the stronger the co-citation relation between the

two publications. In VOSviewer, the strength of co-citation relation will be presented as the thickness of the connection line between spot. In this paper, we can choose to visualize the co-citation of author instead of publication, in that it can show us the impact of author within the research field. Since recursive citation relation is not able to show in the Scopus Method, we manually select the impact of Serra X. to represents the impact of dataset-creating publications of CompMusic Project.

Apart from citation-based bibliometric networks, networks of keywords co-occurrences have been studied extensively[24]. The number of co-occurrences of two keywords is the number of publications in which both keywords occur together in the title, abstract, or keyword list. Keyword co-occurrence network shows the relation between keyword. In VOSviewer, the thickness of the connection line between spot represents the strength of co-occurrence relation and the size of circles represents the number of occurrences of that keyword.

3 Results

3.1 Some examples of the impact of open accessible datasets

In 2014, R. Caro and X. Serra created a corpus of Jingju[20]. We fetched all Music Information Retrieval (MIR) research on Jingju by the query:

```
Jingju MIR
```

We then filtered out irrelevant publications, got 80 publications published within year 2000-2020 (both inclusive). Result is presented in Figure 1.

In Figure 1, blue line represents the number of domain papers published each year, orange line represents the number of “child paper”, which are publications in this field that “recursively” cited paper[20] (that is, it cited [20], or it cited a paper that cited [20] or so on) each year. We can clearly observe the strong correlation between the publication of Jingju corpus and the prosperity of MIR research on Jingju, as the number of publications skyrocketed after the publication of corpus on 2014, and the majority of new publications are inspired by the corpus, directly or indirectly.

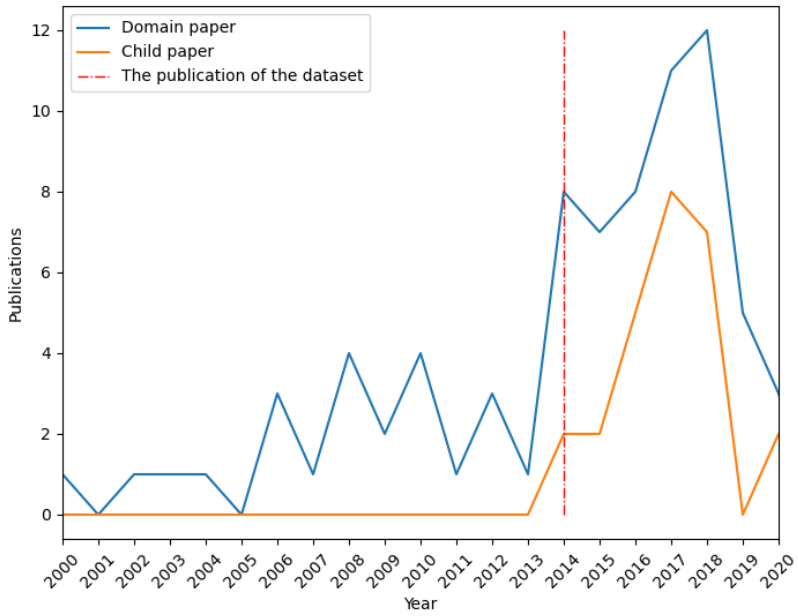


Fig. 1. Publications per year about the MIR work on Jingju (2000-2020)

Using similar approach we analysed the impact of a corpora for music information research in Indian art music, provided by Srinivasamurthy et al[17]. The query used to build domain paper collection is:

```
"Indian art music" MIR
```

the result is presented in Figure. 2.

We can observe the stimulation effect clearly, similar to previous case.

3.2 Indian art music is more popular than Chinese traditional music in the field of MIR

From Fig. 3, we can see that the overall number of publications of Indian art music is much greater than the one of Chinese traditional music, which means the research of Indian art music in the field of MIR is much more popular than Chinese traditional music.

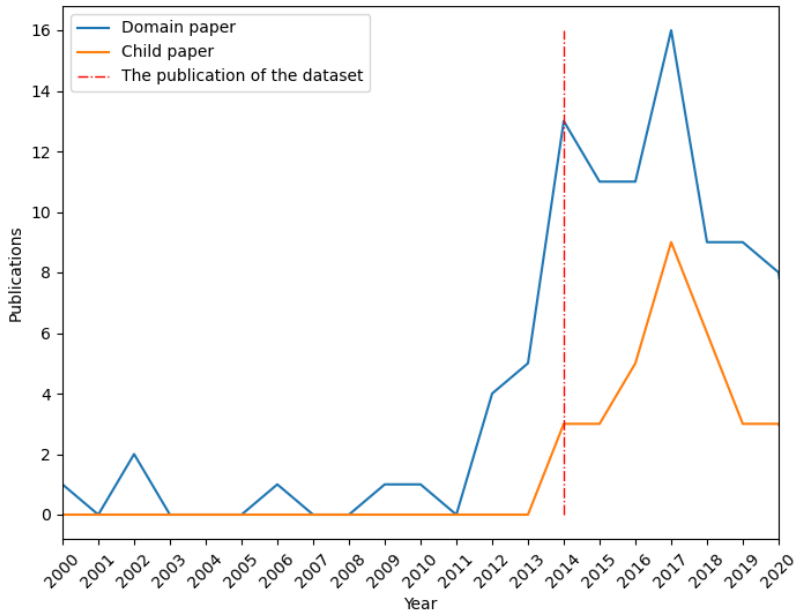


Fig. 2. Publications per year about the MIR work on Indian art music (2000-2020)

3.3 The impact of open access to datasets

From Table. 2, we can know when each dataset was introduced. In 2014, two open datasets (within the same research project: CompMusic Project) of both music traditions were introduced, which correlated to a sudden increase in the popularity in each music tradition （See Figure. 3 eleven publications in Indian art music and five in Chinese traditional music）and each research domain gained the fastest growing speed in the same year. This inference can be supported by the fact that most of these dataset-creating papers have high impact within each music tradition domain, judging from Table. 2. For example, [16] has 23 good citations with a high ratio of 23/29 and this impact is ranked the second in the domain paper collection. In terms of Chinese traditional music, the impact of open dataset paper is much higher than the others. This can be a reason behind the prosperity of the study on Indian art music and Chinese traditional music.

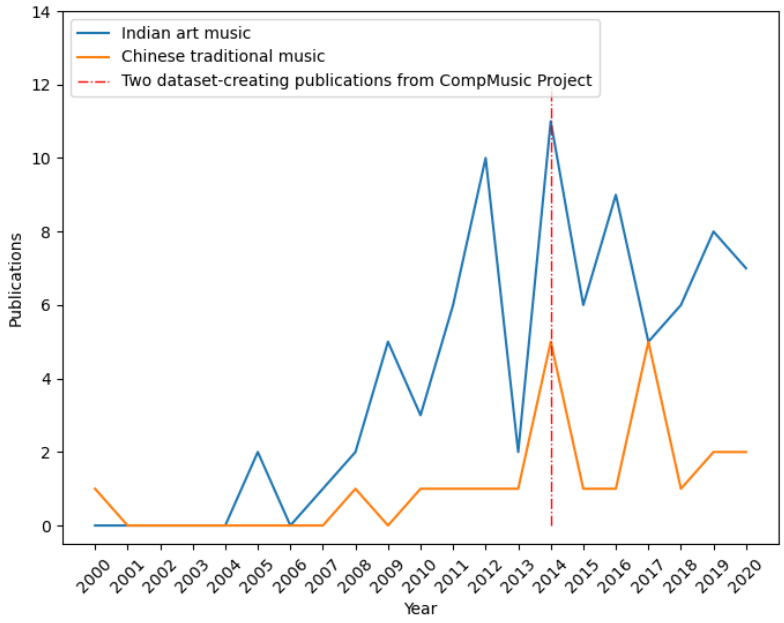


Fig. 3. Publications of Indian and Chinese traditional music in Scopus from 2000-2020

Another intuitive way to demonstrate that the sudden increase in publication number is associated with CompMusic project, is to visualize the co-citation relations bibliometric data in VOSviewer. Figure. 4 and Figure. 5 presents the co-citation network of authors. In the graph, the diameter of each circle represents the number of citations the author has. Serra X. with 50 and 116 citations in the research topic of Chinese traditional music and Indian art music respectively, holds a dominant academic status. In Figure. 4, there are strong links projected from Serra, X. and Indian researcher Rao, P. (Top researcher in Indian art music domain), while there are less thicker connections projected from Chordia, P. (The author of the close dataset NICM2008 in [18]), which means more researches are based on the work of Serra X. and Rao, P. rather than the one of Chordia, P. You can see denser and thicker links in the left side of the network whereas Chordia, P. was left "alone" in the corner. In terms of the authors of close dataset in

Chinese traditional music domain (Xiaojing Liang, Zijin Li et al.), they are not even noticeable in the Figure. 5.

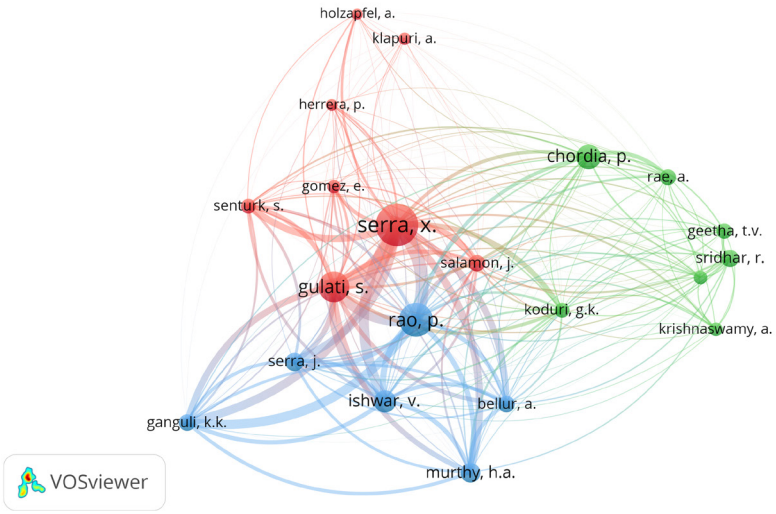


Fig. 4. Co-citation network of author in Indian art music

4 Conclusion and Discussion

In this paper, we applied a bibliometric analysis method to study the impact of openness of dataset to the corresponding field. Through doing co-relation analysis and visualization, we tried to prove that open access to dataset can contribute to the popularity of the research field of Indian art music and Chinese traditional music. Without open access to dataset, the overall popularity of Chinese traditional music is less than the one of Indian art music. Scholars like Serra, X. who led a open research project would have stronger impact than scholars who do not make their research dataset public. In a word, high accessibility of music dataset can lead the research field to high prosperity.

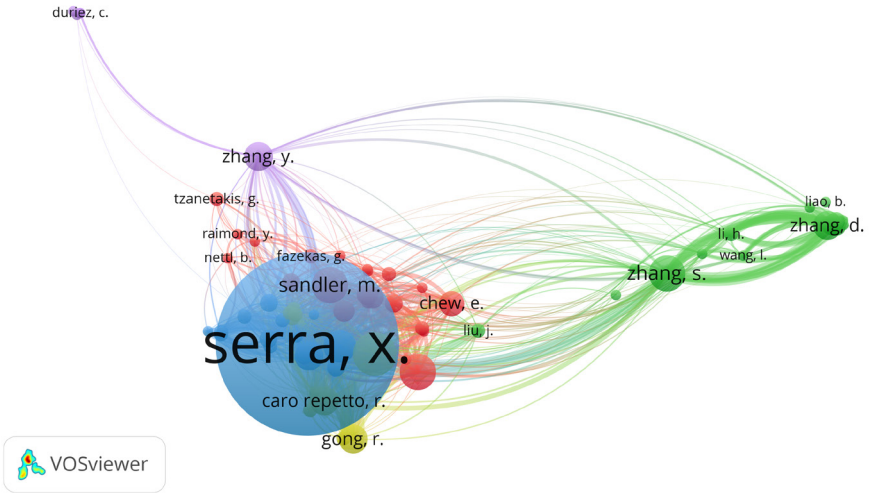


Fig. 5. Co-citation network of author in Chinese traditional music

However, during the research process, we encountered the following problems and we like to point them out:

4.1 Bias in query string

When searching domain papers in Indian art music and Chinese traditional music in Scopus, we discovered that the overall number of publications is small compared to other bibliometric analysis research in literature, which means although we discovered a huge difference in the numbers of retrieved publications, e.g. the domain papers of Indian art music are almost four times bigger than the one of Chinese traditional music, it is still not convincing enough to demonstrate that Indian art music is more popular than Chinese traditional music in the field of MIR. There are several reasons: 1) We are searching papers from global databases, which means mainly English publications are considered. We didn't search regional journal thoroughly. For example, most Chinese publications are not included in Scopus and Google Scholar database. 2) The quality of the open-accessed dataset can affect its contribution to the popularity of the research field. Imagine com-

paring two datasets, one contains raw audio files of various instruments and sufficient annotations while the other mainly contains audio files of singing voice and symbolic data e.g. machine readable scores, it is easy to expect more researches to be carried out based on the former dataset. Actually this is the exact situation when comparing Indian art music dataset and Jingju dataset within CompMusic Project (See Figure. 6, Indian Classical Music is the only music tradition shown in the keywords' network of computational musicology field). In this meta-research project, we had to choose Jingju dataset as a representative of Chinese traditional music dataset (thus as a query string) since it is the most well-known open-accessed Chinese music dataset though it is not an ideal dataset to make comparison to Indian art music dataset of the CompMusic project.

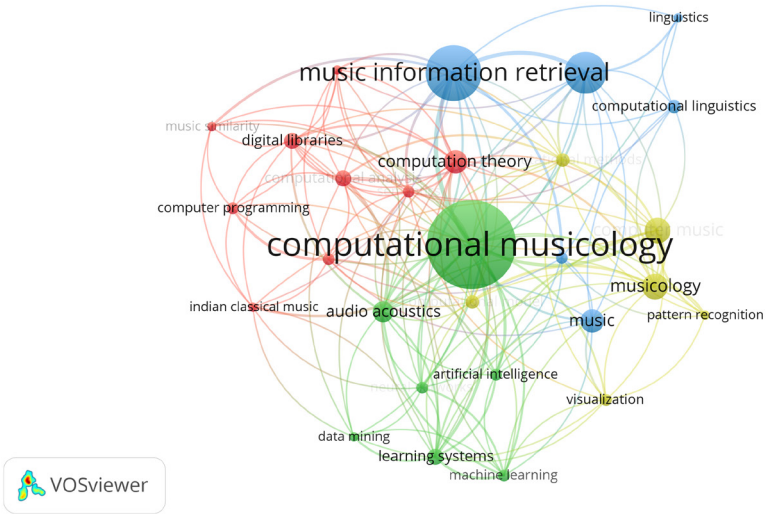


Fig. 6. Co-occurrence analysis of keywords on the retrieved publications in Scopus, query string = "Computational Musicology"

4.2 Definition of research field

How to define a research field? If we choose MIR as the research field, it would be trivial to study the impact of a collection of papers from a minor

research topic (e.g. Computational Musicology) to the whole MIR field. In this paper, we struggled defining the "research field". In the end we searched for a collection of paper which contains one keyword from "music tradition" keywords and one from "domain" keywords. By doing this, the retrieved publications are highly relevant but not plenty enough. Please also note that the different query string are used when searching in Google Scholar and Scopus, which means the results of the two quantitative methods are not comparable with each other, in that different "domain paper" are retrieved based on different query string and different database.

4.3 "Weak" correlation study

Because of the small number of data samples, it is hard to make an association between the publish of a paper and the developing trend of a research field. There may exist some correlations, but logically we can not prove that they are cause and effect relationship. This is the most questionable part in our research. What we can do, is to do our best to make good explanations of the phenomenon.

4.4 The choices of data sources

Google Scholar is a commercial search engine and their data is not public-accessible. Although we used crawler to gather data from Google Scholar, it is still inefficient because the working of our crawler is regularly interrupted by the anti-crawler strategy of Google Scholar. We noticed Microsoft Academic Graph (MAG) during our work, they offered a heterogeneous graph, where the nodes and the edges represent the entities engaging in scholarly communications and the relationships among them, respectively[27]. That's exactly what we need to further draw our conclusion, but due to the time limitation, we were not able to utilize MAG in this project.

5 Appendix

All because of the love for our national music! As a team of three Chinese students, nothing is better than carrying a research project about Chinese

traditional music and shed some lights on the research field. By the way, the current status of the openness of Chinese Traditional music dataset is getting better. In November 2021, after interviewing Xiaojing Liang, an author of a previously close music dataset Chinese Traditional Instrument Sound Database (CTIS), we are glad to find that a collection of Chinese traditional music dataset became accessible to the public, namely CTIS, Midi-wav Bi-directional Database of Pop Music and Multi-functional Music Database for MIR Research (CCMusic). One can find more information at `https://ccmusic-database.github.io/en/overview.html`. Maybe it is a good idea to study the impact of this collection of music dataset five years later!

We utilized Publish or Perish 8 in our data analysis procedure. Our script based on the scholar project forked by fjxmlzn. The python scripts used in data gathering and analysis are accessible at `https://github.com/seeker-Liu/ResearchMethod`

References

[1] A. Koh, "Music for ai reports: Dual prospects in music production," 2018.

[2] T. J. O'shea and N. West, "Radio machine learning dataset generation with gnu radio," in *Proceedings of the GNU Radio Conference*, vol. 1, 2016.

[3] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[4] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis) contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, p. 100 336, 2021.

[5] K. Antelman, "Do open-access articles have a greater research impact?" *College & research libraries*, vol. 65, no. 5, pp. 372–382, 2004.

[6] H. Piwowar, J. Priem, V. Larivière, *et al.*, "The state of oa: A large-scale analysis of the prevalence and impact of open access articles," *PeerJ*, vol. 6, e4375, 2018.

[7] Y. Gargouri, C. Hajjem, V. Larivière, *et al.*, "Self-selected or mandated, open access increases citation impact for higher quality research," *PloS one*, vol. 5, no. 10, e13636, 2010.

[8] C. Hajjem, S. Harnad, and Y. Gingras, "Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact," *arXiv preprint cs/0606079*, 2006.

[9] G. Eysenbach, "Citation advantage of open access articles," *PLoS biology*, vol. 4, no. 5, e157, 2006.

[10] P. M. Davis and W. H. Walters, "The impact of free access to the scientific literature: A review of recent research," *Journal of the Medical Library Association: JMLA*, vol. 99, no. 3, p. 208, 2011.

[11] B.-C. Björk and D. Solomon, "Open access versus subscription journals: A comparison of scientific impact," *BMC medicine*, vol. 10, no. 1, pp. 1–10, 2012.

[12] I. D. Craig, A. M. Plume, M. E. McVeigh, J. Pringle, and M. Amin, "Do open access articles have greater citation impact?: A critical review of the literature," *Journal of Informetrics*, vol. 1, no. 3, pp. 239–248, 2007.

[13] W. Chen, J. Keast, J. Moody, *et al.*, "Data usage in mir: History & future recommendations," *20th International Society for Music Information Retrieval Conference*, 2019.

[14] J. Wang, T. Zheng, Q. Wang, B. Xu, and L. Wang, "A bibliometric review of research trends on bioelectrochemical systems," *Current Science*, vol. 109, pp. 2204–2211, 2015.

[15] E. Abdeljaoued, M. Brulé, S. Tayibi, *et al.*, "Bibliometric analysis of the evolution of biochar research trends and scientific production," *Clean Technologies and Environmental Policy*, vol. 22, pp. 1967–1997, 2020.

[16] X. Serra, "Creating research corpora for the computational study of music: The case of the compmusic project," 2014, pp. 1–9.

[17] A. Srinivasamurthy, G. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," 2014, pp. 1029–1036.

[18] P. Chordia, M. Godfrey, and A. Rae, "Extending content-based recommendation: The case of indian classical music," 2008, pp. 571–576.

[19] "Google scholar, web of science, and scopus: A systematic comparison of citations in 252 subject categories," *Journal of Informetrics*, vol. 12, no. 4, pp. 1160–1177, 2018, ISSN: 1751-1577.

[20] R. C. Repetto and X. Serra, "Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis," in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014, pp. 313–318.

[21] X. Liang, L. Zijin, J. Liu, W. Li, J. Zhu, and B. Han, "Constructing a multimedia chinese musical instrument database," in Jan. 2019, pp. 53–60, ISBN: 978-981-13-8706-7.

[22] 李子晋 (LI Zijin), 于帅 (YU Shuai), 肖畅 (XIAO Chang), 耿瑜曼 (GENG Yuman), 钱文琪 (QIAN Wenqi), 高永伟 (GAO Yongwei), 李伟 (LI Wei), "CCMusic: 用于 MIR 研究的中国音乐数据库建设 CCMusic Database: Construction of Chinese Music Database for MIR Research," *复旦学报 (自然科学版) Journal of Fudan University(Natural Science)*, vol. 58, no. 3, pp. 351–357, 2019.

[23] 李圣辰 (LI Shengchen) and 吴雨松 (WU Yusong), "1 个中国古琴曲的符号化音乐数据集介绍及其应用实例 An Introduction to a Symbolic Music Dataset of Chinese Guqin Pieces and Its Application Example," *复旦学报 (自然科学版) Journal of Fudan University(Natural Science)*, vol. 59, no. 3, pp. 276–285, 2020.

[24] N. J. van Eck and L. Waltman, "Software survey: Vosviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010.

[25] A. M. Aragón, "A measure for the impact of research," *Scientific reports*, vol. 3, no. 1, pp. 1–5, 2013.

[26] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[27] K. Wang, I. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: When experts are not enough," *Quantitative Science Studies*, vol. 1, no. 1, pp. 396–413, 2020.

Evaluation of countries based on their use of Open Science in medical research: a comparison with reference rankings

Eduard Alcobé Garcia
Miguel Augusto Silva Fuentes
Aaron Verdaguer Gonzalez

Master in Intelligent Interactive Systems
eduard.alcobe01@estudiant.upf.edu
miguelaugusto.silva01@estudiant.upf.edu
aaron.verdaguer01@estudiant.upf.edu

Abstract. Nowadays there are many reliable indexes to rank countries based on the quality of their open science policies, health care, or medical research power. However, are they useful to evaluate open science use in the medical field per country? In this paper, five of them have been chosen to compare them with a newly developed metric which ranks countries according to the use of open science in the medical field. The creation of a new ranking based on this new metric is thought to correlate open science policies with medical research power, and to study that, twenty different countries across the world are selected. Furthermore, it intends to evaluate the use of open science in medical research in these twenty countries.

Keywords: Health care quality, open science, open data, rankings, medical research, countries, indicators.

1 Introduction

Health care quality is one of the most important factors in order to define and quantify quality of life. For this reason, many efforts have been made by countries to improve health care [1]. In this aspect, it is believed that the correct use of open science and open health databases may be a major breakthrough to approach the Culture of Health [2]. Theoretically, sharing medical data worldwide would provide researchers with a huge amount of data to analyse and compare, as well as the possibility of having complete databases to reach better treatments and faster diagnosis. Additionally, it may be essential to fight dangerous epidemics humanity has faced [3]. Trusting in the use of open science would possibly lead to a significant process in terms of health care quality, just because all the information would be available through Health Database Organizations [4].

On one hand, some countries have recently relied on open science and the number of publications has been increasing. Specifically, openness has experienced a quick rise basically thanks to the adoption of downloading options. Furthermore, licenses

regulation, provision of data, and machine-readable file formats have contributed as well. Meanwhile, coverage elements have been improving gradually [5][6]. On the other hand, health quality has experienced some changes as well in recent years thanks to medical research [7].

The aim of this work is to qualify the use of open science in the field of medicine in different countries. Another goal is to determine if current available rankings in open science, in health quality, or in research collaboration per country are representative from the point of view of open access in medical research. The open science rankings used are the Open Data Inventory (ODIN) ranking [5] and the study done by Open Science Monitor (OSM) [6]. The health quality ranking used is the STC health index (STC). Finally, two research rankings are compared as well, the Nature index in life science (NI) [8] and the SJR index in medicine (SJR) [9]. In order to evaluate the impact of open science in medicine and develop a ranking, different indicators are taken into account to develop the most accurate metric as possible. Those are the number of publishers per country in the most important medical open access journals according to Scopus, the number of publications per country available in the Directory of Open Access Journal (DOAJ) in the medicine field [10] and the number of repositories per country and funders' policies depending on the country available in Re3data [11].

For this reason, in this study, it is wondered whether available rankings in open science, health, or medical research in general per country are representative of the use of open science in medicine. Interestingly, some countries may be ranked quite high regarding open science policies, however, their use of it in medicine may not be prominent. In addition, from a more theoretical perspective, it is tried to discern if the use of open science in a country directly affects health care.

Improving health care has always been one of the main objectives of human civilization. Nowadays open science brings a feasible way to make a step forward as research in medicine may increase. Moreover, comparing countries' open science policies may be possible to evaluate the importance of open science as well as their quality. Providing new knowledge and a new metric tool; through the use of available data, rankings, and indicators about open science applied in medicine; would help to evaluate the confidence of some open science rankings as well as of open science applied in medicine by country. Perhaps, it may even help to convince sceptical medical researchers to use open access journals and public databases.

2 Research methodology

For the development of this research, the project focuses on indicators of great importance to evaluate the quality of research within open science. The objective is to rank 20 selected countries in regard to different indicators. Countries have been chosen according to their localization, to ensure that the widest possible spectrum of regions is evaluated. The studied countries are the most powerful ones according to the rankings in open science and in research used in this project.

The first indicator is the number of publishers per country in the most important open access journals from health across the world according to CiteScore [12]: Molecular Cancer, The Lancet Global Health, Annals of the Rheumatic Diseases, and The Lancet Public Health. It has been proven to be a very important variable to determine the importance of journals due to its multiple metrics (citations by total of publications given a time). Research relies on the context in which it will be analysed [13], in this case, the metrics for the medical field.

The next chosen indicator is the number of open access directories per country available in the Directory of Open Access Journals (DOAJ) [10]. This tool is used to identify which are the countries with the largest number of open access medical directories to achieve the goal of being as objective as possible and developing a representative metric to rank countries according to open science in medical research.

The third indicator used in the current research is the number of open data repositories in Re3data. Re3data is a global registry of research data repositories that covers research data repositories from different academic disciplines [14]. The Lisbon Council, ESADE, and Elsevier analysed Re3data at the subject level [6], so it can be easily evaluated on the quality of this research at the health data level and take into account the data at the country level too. This analysis gives more information, as the trends in open science and drivers can help to facilitate the adoption of it.

Based on these indicators, the objective is to create a weighted average and see if there is a correlation between the different variables and how reliable is the rating of each of these with regard to the majority, this may be helpful to reach a conclusion in terms of open data use in the medicine area. These three indicators are used in order to represent the power of a country in open science applied in medical research as the number of publications and repositories are clear indicators to quantify this power mentioned. The formula used is the following:

$$Score = \frac{\sum \%\ of\ total\ publications}{number\ of\ journals} + \%\ DOAJ + \%\ Re3data$$

Finally, an exploratory analysis is carried out on the resulting ranking generated based on the new score and the other rankings. The analysis takes into account 5 rankings. The first one is the ODIN Score [5], created by Open Watch, which presents us with another ranking with multiple metrics on the development of open data (such as data available last five years, terms of use, machine readable, etc.) within different categories. Also, another open science ranking used is the open science monitor [11] part of the Strategy 2020-2024 of the European Union, a big analysis based on different sources. The third is the STC Health Index [7], which has different very important metrics, such as health life expectancy at birth, skilled health professional density, etc. The fourth one is the Nature Index [8], a database of author affiliations and institutional relationships, in this case, only the data available in life science is going to be used. Finally, the Scimago Journal & Country Rank [9]. This country ranking is based on Scopus data and has a lot of interesting metrics that give us different points of view, for

instance the total number of citable documents, the number of references, or the average citation by document.

3 Results

Figure 1 depicts the ranking of the 20 chosen countries based on the created metric and Table 1 represents the contribution by percentage of each country for each of the studied variables. As mentioned previously the criteria to follow in order to create such a metric is the weighted average of the number of publications, the number of journals in DOAJ, and the number of data repositories in re3data. More importance has been given to the last 2 variables thus many countries do not publish in the top 4 open access journals in medicine, but they have very competent ones in their country. In figure 1 it can be observed how open access medical research quality based on the new metric is not strongly correlated to any of the other variables in the Appendix.

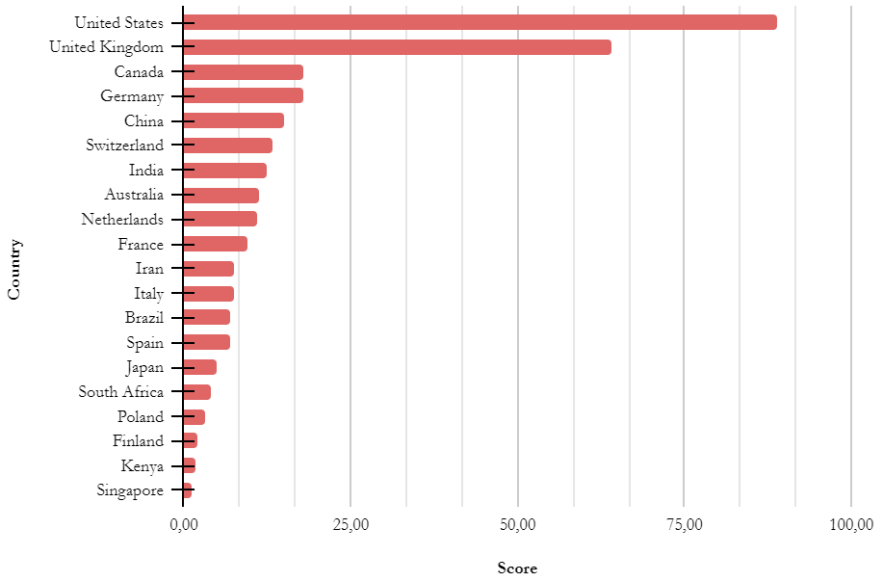


Figure 1. Top 20 countries in Open Science in the health field based upon the created score.

The appendix includes the rankings used to conduct the exploratory analysis. As mentioned previously the first one is the ODIN ranking (Figure A1). This ranking does not reflect a complete similarity with the obtained ranking (Figure 1), but there are some countries that generate correlation like China, Germany, Canada, and Switzerland. These countries remain in the top 10, but not always in the same order. The other open science ranking (Figure A2) just provides information of 14 out of the

20 studied countries (Australia, Iran, Brazil, South Africa, Kenya, and Singapore are not included). Even though there are 6 countries missing, there are many misclassifications as well. The third ranking is about the STC Health Index (Figure A3), as the results show, there is no relationship between this ranking and the one proposed in this paper. The fourth one from Nature (Figure A4) also shares some similarities within the ten first countries, which are in a different order but remain within the top 10. About the last ranking, the SJR Index (Figure A5) it can be seen that the behaviour of the top 10 countries is very similar to that of the previous ranking. The 20 studied countries are ranked from 1st to 20th (only to 14th in the open science monitor ranking shown in figure A2). For the purpose of evaluating how similar these 5 rankings are to the new created, the number of misclassifications is calculated and presented in table 2. As shown in Table 2, the Nature Index and the SJR Index rankings are the most similar to each other and to the one presented in this work. The number of misclassifications per number of evaluated countries is significantly lower than in the other rankings.

	Journals				Repositories			
Country	**Molecular Cancer**	**The Lancet Global Health**	**Annals of the Rheumatic Diseases**	**The Lancet Public Health**	**DOAJ**	**Re3data**	**SCORE**	**Ranking**
United States	35,39	40,24	16,60	24,31	8,56	51,08	88,77	1
United Kingdom	5,67	31,71	27,55	38,27	23,44	14,92	64,16	2
Canada	3,81	7,08	4,29	9,99	1,22	10,31	17,82	3
Germany	7,58	3,39	9,55	3,49	1,17	10,62	17,79	4
China	35,61	5,29	2,03	5,17	1,12	2,00	15,15	5
Switzerland	1,53	11,89	3,75	3,73	4,16	3,85	13,23	6
India	2,61	9,17	0,53	1,44	6,79	2,31	12,54	7
Australia	2,87	8,38	3,54	11,91	1,01	3,54	11,23	8
Netherlands	1,68	4,36	13,15	4,69	2,99	2,15	11,11	9
France	3,70	4,62	8,74	8,42	0,31	2,77	9,45	10
Iran	0,30	0,97	0,06	0,60	7,08	0,00	7,56	11
Italy	5,49	1,79	6,79	2,89	1,48	1,69	7,41	12
Brazil	0,90	3,91	0,92	1,08	5,10	0,15	6,96	13

Spain	2,84	2,46	5,28	2,53	2,55	1,08	6,90	14
Japan	4,03	1,01	3,84	0,72	0,55	1,85	4,79	15
South Africa	0,04	9,31	0,45	1,32	0,73	0,46	3,97	16
Poland	0,41	0,41	0,86	0,60	2,55	0,15	3,27	17
Finland	0,56	0,67	2,17	2,65	0,18	0,31	2,00	18
Kenya	0,04	4,40	0,03	0,48	0,08	0,31	1,62	19
Singapore	1,01	1,27	0,24	0,72	0,18	0,15	1,15	20

Table 1. Percentage of publications in each of the selected journals, number of journals in the DOAJ, and number of open data repositories for the selected countries.

ODIN misclassification	**OSM misclassification**	**STC misclassification**	**NI misclassification**	**SJR misclassification**
3,20	2,57	3,05	1,45	1,55

Table 2. Number of misclassifications compared with the created ranking and weighted with the number of countries evaluated in each ranking.

4 Conclusions

The achievement of a new ranking to classify countries' use of open science in the medical field (Figure 1) shows interesting results. Clearly, the United States of America and the United Kingdom dominate, probably due to their economic power as well as the use of English as natives. Surprisingly, technological countries such as Japan, Finland, or even Poland are ranked quite low. A possible reason for this is that they are not leading countries in medical research although their sanitary system may be professional and their open science policies strong. The Japan case is quite notable as it is well ranked in all the evaluated rankings, however, in the created one a low score is achieved. Globally, the ranking shows the dominance of the rich European countries as well as the most powerful countries of the world. Nevertheless, India, Iran, and Brazil, which may not be called the richest, make a significant contribution to open science in the medical field. These countries have a large number of inhabitants; therefore, they have more possibilities to make an impact than other countries. It is interesting to

highlight the India case as it is ranked 7th in the presented ranking while in all other rankings it does not achieve such an acceptable position.

According to the developed metric, existent rankings in open science per country (Figure A1 and Figure A2) do not seem accurate enough to describe the use of open science in the medicine field. Lots of indicators are taken into consideration and a complex methodology is followed in those rankings, therefore the results may be distant from those found in the analysis presented. Also, the health ranking (Figure A3) used is not in tune with the developed ranking. Hence, no direct effect of open science use in medical research is found, the most reasonable explanation is that lots of variables have a higher impact in health care than the use of open science for research in that field. Nevertheless, interestingly the resulting ranking is quite in accordance with rankings sorting countries by their contribution in medical research (Figure A4 and Figure A5).

Taking this into consideration, it can be concluded that the most important countries in research are the ones dominating the medicine investigation using open science, which may not be surprising. In addition, some countries with great open science policies such as Singapore or Finland suffer from a lack of research compared with other countries like the United States of America or the United Kingdom. Therefore, it can be highlighted that open science policies and the collaboration of open science (in this case in the medical research) are not going together, consequently, more improvement may be done in that area. Leading countries in research do not have the best open science policies whereas the countries with better open science do not have enough research power to make a difference. Nevertheless, some countries such as Germany or the Netherlands seem to have an equilibrium between open science policy, health, and medical research because they are ranked similarly in all rankings. Perhaps, it is the moment to follow their example and combine both (open science policy and medical research power) in order to take a clear advantage of open science use at the sacrifice of traditional hermetic research. The possibility is there as well as the benefits, the only thing missing is the desire and the initiative to do it.

References

1. Bullinger, A. C., Rass, M., Adamczyk, S., Moeslein, K. M., & Sohn, S. (2012). Open innovation in health care: Analysis of an open health platform. *Health Policy*, *105*(2–3), 165–175. https://doi.org/10.1016/j.healthpol.2012.02.009
2. Rowhani-Farid, A. (2018). *TOWARDS A CULTURE OF OPEN SCIENCE AND DATA SHARING IN HEALTH AND MEDICAL RESEARCH*.
3. D'Agostino, M., Samuel, N. O., Sarol, M. J., de Cosio, F. G., Marti, M., Luo, T., Brooks, I., & Espinal, M. (2018). Open data and public health. In *Revista Panamericana de Salud Publica/Pan American Journal of Public Health* (Vol. 42). Pan American Health Organization. https://doi.org/10.26633/rpsp.2018.66

4. Molla S. Donaldson and Kathleen N. Lohr., & Committee on Regional Health Data Networks, I. of Medicine. (1994). *Health Data in the Information Age : Use, Disclosure, and Privacy.* National Academies Press.
5. Open Data Inventory. (2021). *Annual report of 2020/21*.
6. ESADE, & The Lisbon Council. (2019). *OPEN SCIENCE MONITOR STUDY ON OPEN SCIENCE: MONITORING TRENDS AND DRIVERS*.
7. Hudson International Group. (2021). *The 2021 STC HEALTH INDEX.*
8. A guide to the Nature Index. (2017). Nature, 548(7666). https://doi.org/10.1038/548S32a
9. SCImago, (n.d.). SJR — SCImago Journal & Country Rank [Portal]. Retrieved Date you Retrieve, from http://www.scimagojr.com
10. Morrison, H. (2017). Directory of Open Access Journals (DOAJ). *The Charleston Advisor*, *18*(3). https://doi.org/10.5260/chara.18.3.25
11. European Union, European Commission. *Facts and Figures for open research data.* Retrieved from https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor/facts-and-figures-open-research-data_en#funderspolicies
12.van Noorden, R. (2016). Controversial impact factor gets a heavyweight rival. In *Nature* (Vol. 540, Issue 7633). https://doi.org/10.1038/nature.2016.21131
13. Colledge, L., James, C., Azoulay, N., Meester, W., & Plume, A. (2017). CiteScore metrics are suitable to address different situations – A case study. *European Science Editing*, *43*(2). https://doi.org/10.20316/ESE.2017.43.003
14.Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H. J., Gundlach, J., Schirmbacher, P., & Dierolf, U. (2013). Making research data repositories visible: The re3data.org registry. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0078080

Appendix

Source title	CiteScore	Highest percentile	2017-20 Citations	2017-20 Documents	% Cited	SNIP	SJR	Publisher
Molecular Cancer	34,3	99,0% 2/167 Molecular Medicine	21863	637	96	4.213	7.274	Springer Nature
The Lancet Global Health	32,1	99,0% 4/793 General Medicine	13014	406	96	10.022	7,97	Elsevier
Annals of the Rheumatic Diseases	28,7	99,0% 1/56 Rheumatology	24277	845	94	4.294	6.333	BMJ Publishing Group
The Lancet Public Health	28,7	99,0% 2/526 Public Health, Environmental and Occupational Health	5338	186	95	7.171	7.226	Elsevier

Table A1. Selected journals for the analysis of the number of publications with their CiteScore and other indicators which validates them.

	Journals				Repositories	
Country	Molecular Cancer	The Lancet Global Health	Annals of the Rheumatic Diseases	The Lancet Public Health	DOAJ	RE3DATA
United States	948	1080	2718	202	329	332
United Kingdom	152	851	4511	318	901	97
Germany	203	91	1564	29	45	69
Canada	102	190	702	83	47	67
India	70	246	87	12	261	15
Switzerland	41	319	614	31	160	25
China	954	142	333	43	43	13
Netherlands	45	117	2152	39	115	14
Australia	77	225	580	99	39	23
France	99	124	1431	70	12	18
Spain	76	66	865	21	98	7
Italy	147	48	1112	24	57	11
Japan	108	27	628	6	21	12
South Africa	1	250	74	11	28	3
Kenya	1	118	5	4	3	2
TOTAL	2679	2684	16371	831	3844	650

Table A2. Number of publications for the selected journals, number of open journals (DOAJ), data repositories (Re3data) for the selected countries, and global contributions to each journal or repository.

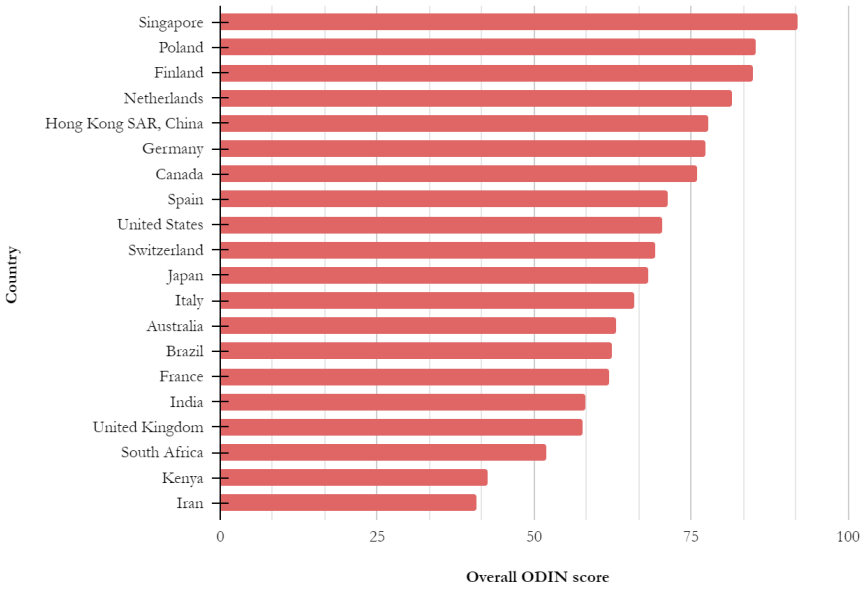


Figure A1. Rank of studied countries based on the overall ODIN score.

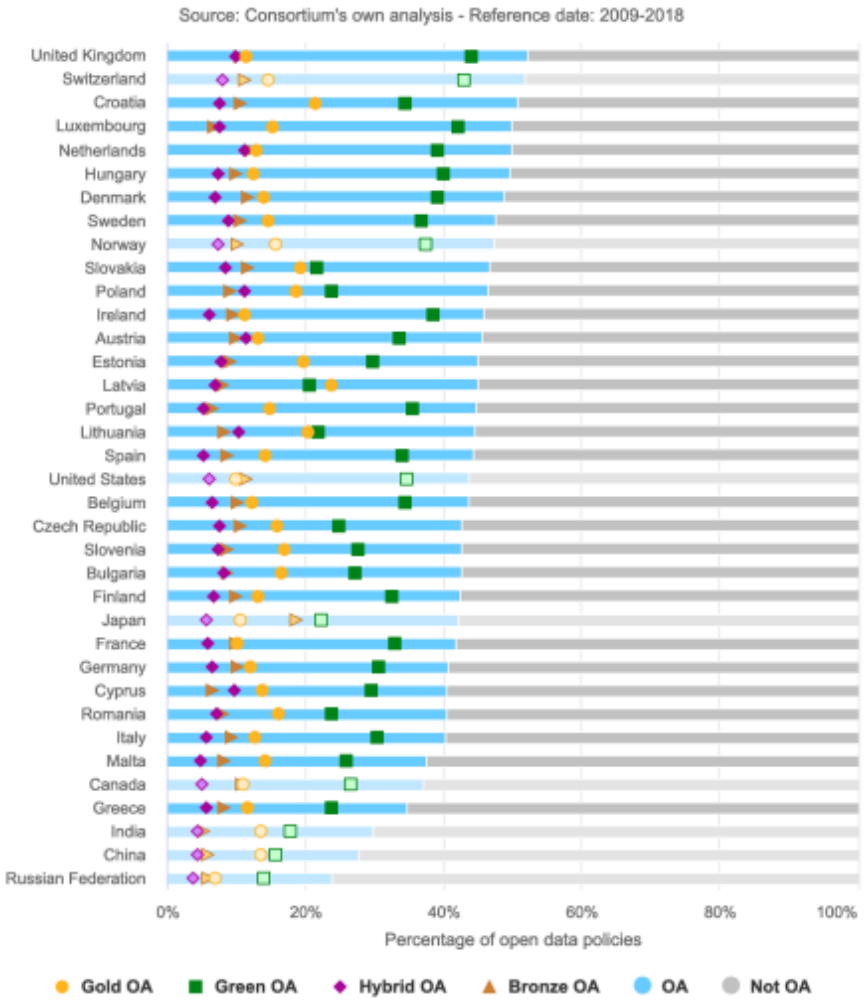


Figure A2. Ranking of countries based on the percentage of open data policies [6].

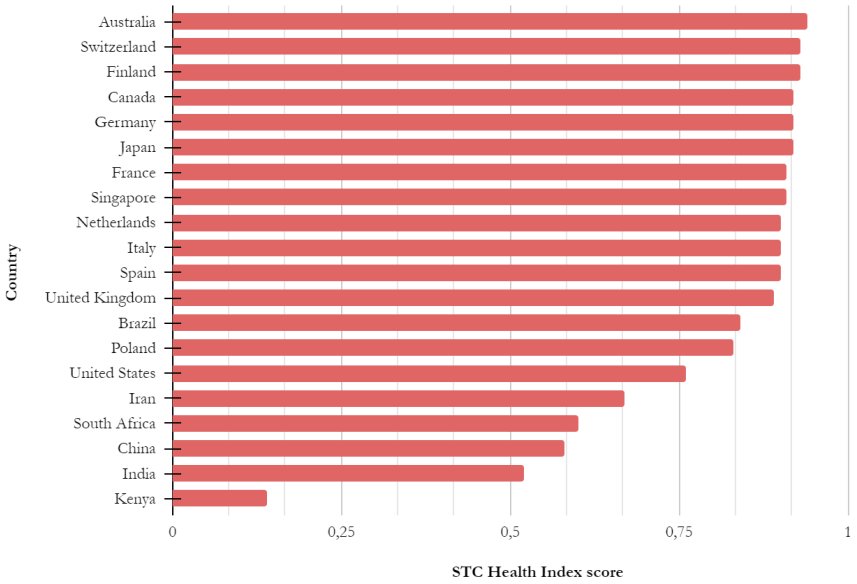


Figure A3. Rank of studied countries based on the STC Health Index.

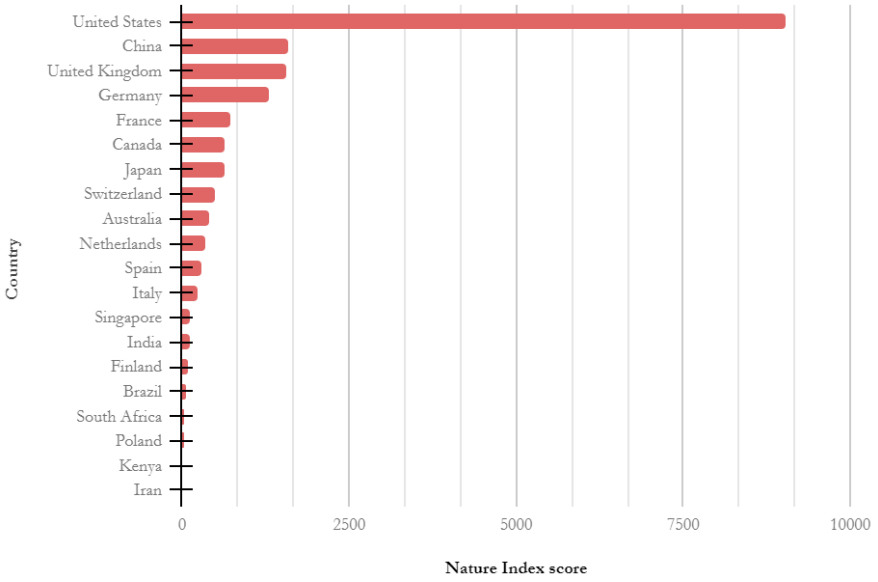


Figure A4. Rank of studied countries based on the Nature Index.

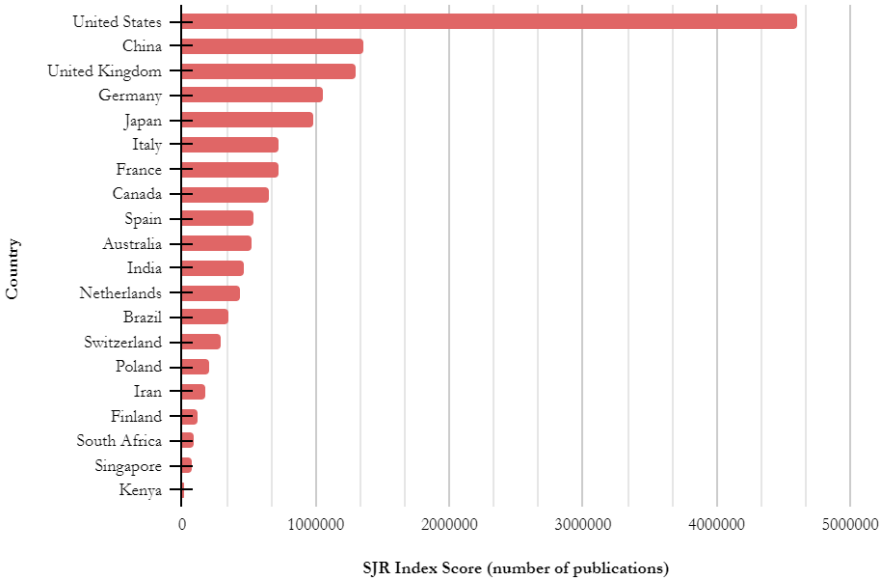


Figure A5. Rank of studied countries based on the SJR Index.

An Evaluation of Readability Metrics for Scientific Writing

Davide Locatelli, Ilse Meijer, Stephanie Rodriguez

Master in Intelligent Interactive Systems
davide.locatelli01@estudiant.upf.edu
ilse.meijer01@estudiant.upf.edu
stephanie.rodriguez01@estudiant.upf.edu

Abstract. A readability score indicates how difficult a text is and can be used to match readers with material based on their language proficiency. While both analytic formulas and deep learning algorithms have been proposed for automatic readability assessment, state-of-the-art results are measured on educational text aimed at young readers. Here, we investigate whether these techniques generalize well to readability assessment of scientific writing. We assess both popular analytic approaches including the Gunning Fog, the Dale-Chall, and the Flesch Reading Ease indexes as well as deep learning algorithms based on BERT. Our results show that none of these methods can produce accurate results and that more research around readability of scientific writing is needed. We conclude that an accurate analytical solution needs to make use of more reliable linguistic features, and on the other hand an accurate deep learning classifier requires larger and more diverse datasets.

Keywords: Meta-Research, Readability, Scientific Writing, Readability formulas, BERT, Natural Language Processing

1 Introduction

Readability is defined as the difficulty to understand a written text [1, 2]. Automatic readability assessment (ARA) is the task of predicting a readability score to a given text. Various techniques have been proposed for ARA. On the one hand, over two hundred formulas have been proposed as an analytical solution [3]. These approaches generally base their calculations on lexical and syntactic features such as the number of difficult words and the length of sentences. On the other hand, deep learning algorithms have been used to provide predictive functions based on auto-generated features learned from data, achieving state-of-the-art performance on popular readability benchmarks [4].

Noticeably, the vast majority of the studies that use the analytical formulas have focused on children and high school texts [3, 4]. Moreover, the promising results of machine learning have been obtained by testing on datasets compiled using educational textbooks as the primary source, such as WeeBit [5] and OneStopEnglish [6]. Hence, there remains an open question as to how accurately these methodologies can predict readability scores of text meant for audiences other than young readers. In this paper we evaluate the performance of the above

mentioned methodologies by testing them on scientific writing. Several studies have demonstrated that scientific papers are getting progressively harder to read [2,7,8]. As science becomes more detailed and specialized, the expertise required to understand papers has drastically escalated [2]. As a result, knowledge circulates less freely between the disciplines, which become increasingly isolated.

Arguably, this has significant consequences for both young and senior researchers. Young researchers are faced with the additional challenge of understanding highly technical language, while having to simultaneously advance the grasp of the field they wish to enter. Senior researchers are faced with a higher cost of exploring new disciplines for similar reasons, hence getting more confined to their current specialization. A readability score can thus provide researchers with an idea of the difficulty of a paper prior to reading. This would allow them to select more accessible material at first, and proceed with more difficult research in due course, thus facilitating them as they approach a new field.

In this work we focus on papers sampled from Natural Language Processing conference proceedings, using crowd-sourced readability scores as a test set for the most popular readability formulas and deep learning algorithms. Our research questions are as follows:

1. Can traditional formulas be reliably used to assess the readability of scientific texts?
2. Can state-of-the-art machine learning algorithms trained on standard ARA datasets generalize to scientific writing texts?

We hypothesize that deep learning methods produce more accurate scores for our test set because they are not designed to take into account handcrafted features, thus having the potential to capture more general patterns. We believe that finding an accurate ARA predictor for scientific texts would have a positive impact for science, as it would enable readers to select research papers based on their language proficiency and advance their knowledge of difficult topics with more ease.

2 Research methodology

2.1 Corpus

The corpus used in this study was generated by sampling papers from these Natural Language Processing conferences: the *North American Chapter of the Association for Computational Linguistics* (NAACL), the *International Conference on Computational Linguistics* (COLING), and the *Conference on Natural Language Learning* (CoNLL). Our choice of conferences was based on their h5-index score, which is the h-index for articles published in the last five years: it is the largest number h such that at least h published in the last five years have at least h citations each. We aimed to select influential conferences, while keeping the corpus diverse enough. NAACL and COLING are both in the top-5, while

CoNLL is tenth in the ranking so these conferences represented an ideal choice for our test set.[1]

A total of 15 papers were selected. We ensured that each paper had a unique first author and we then extracted three sections of different lengths up to 100 words for each paper, obtaining 45 short excerpts. This was done in order to facilitate the labeling effort by having human annotators be required to read short excerpts instead of the full paper. We controlled excerpt selection for topics in order to obtain sections with different writing styles but with similar contents.

2.2 Participants

We recruited 24 participants among junior researchers in a number of different fields, including Computer Science, Natural Language Processing, Mathematics, Natural and Social Sciences. We recorded information about their background and we report it in Figure 1.

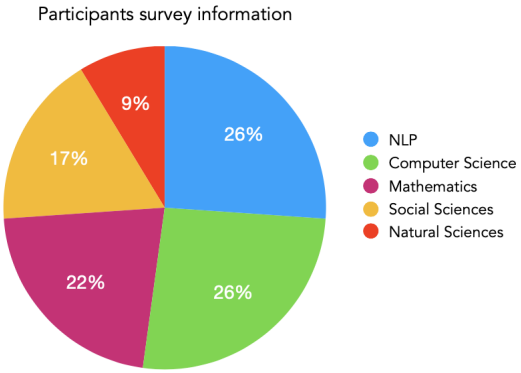


Fig. 1. Distribution of participants by their background. The majority of the annotators had a background in NLP and Computer Science, closely followed by Mathematics and Social Sciences.

2.3 Survey design

In our survey, participants were presented with random excerpts and were asked to categorize each of the texts into the levels of readability: easy, medium and hard. Each annotator had to label at least six texts, and could choose to continue to annotate the entire sample. The gold label was then calculated by taking the majority label for each text.

[1] We used the Google Scholar ranking available at `https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics`.

We considered the difference in backgrounds of our annotators to be an important source of confounding. In order to minimize the impact of this confounding variable, we ensured that the excerpts in our corpus were selected so that they contained enough information to be understood on their own, without access to the rest of the paper or need for prior knowledge.

As a sign of robustness of our survey we report in Table 1 one example of texts that were about similar topics but that were assigned different gold labels. In general we found that this occurred frequently. This trend is arguably a sign that annotators' were more likely to base their decision for a readability label on the linguistic features of the text rather than their proficiency with the topic.

Table 1. An example pair of texts about a similar topic (machine translation) that have been assigned two different gold labels.

Text	Gold label
"In historical linguistics, cognate detection is the task of determining whether sets of words have common etymological roots. Inspired by the comparative method used by human linguists, we develop a system for automated cognate detection that frames the task as an inference problem for a general statistical model consisting of observed data (potentially cognate pairs of words), latent variables (the cognacy status of pairs) and unknown global parameters (which sounds correspond between languages). We then give a specific instance of such a model along with an expectation-maximisation algorithm to infer its parameters."	hard
"Existing approaches to automated cognate detection (ACD) fail to fully capture this idea of dealing with mutual dependence using an iterative method. Some early approaches are not iterative at all, while several more recent methods are iterative to some extent but either only carry out a small fixed number of iterations or use incomplete and ad hoc methods to update sound correspondences based on tentative cognacy judgements. In this paper we design and implement an iterative algorithm that uses the method of expectation maximisation for statistical inference, which is close to historical linguists' method of updating sound correspondences in one iteration based on cognacy judgements from the previous iteration."	intermediate

To ensure high quality gold labels we calculated the inter-annotator agreement between experts and non-experts using the kappa score metric [9,10].[2] The kappa score is 0.28, which is classified by its authors as a fair agreement score. Furthermore we ensured that each excerpt was labeled by at least 8 annotators to ensure that the gold label is of high quality.

[2] The kappa score metric is used to measure the agreement of different annotators by also taking into account the agreement due to chance. It ranges from -1 to 1. Scores lower than zero signal that agreement occurs by chance, while 1 represents complete agreement between annotators.

2.4 Experiments

We tested the following readability formulas on our test corpus:

1. Flesch Reading Ease (FR) [11]
2. Dale-Chall readability formula (DC) [12]
3. Gunning Fog Index (GF) [13]

These are calculated as follows:

$$FR = 206.835 - 1.015 \times ASL - 84.6 \times ASW \tag{1}$$

$$DC = 15.79 \times PHW + 0.0496 \times ASL \tag{2}$$

$$GF = 0.4 \times (ASL + PHW) \tag{3}$$

where ASL is the average sentence length, ASW is the average number of syllables per word, and PHW is the percentage of hard words.

These formulas, like many others, calculate readability by assigning text to school grades. We opted for these particular ones because they include university grades, which is the target group of our research. Moreover, the Flesch Reading Ease and the Dale-Chall are among the most popular formulas by citations. To convert the scores of our methods to the three readability levels we categorize grade levels below college freshmen as easy and above college graduates as hard. The conversion rules are reported in Table 2.

Table 2. Gold label conversion rules for readability formulas

	easy	**intermediate**	**hard**
Flesch Reading Ease score	$50 < x \leq 60$	$30 < x \leq 50$	$x \leq 30$
Dale-Chall readability formula	$x \leq 8.99$	$9 \leq x < 9.9$	$x \geq 10$
Gunning Fog Index	$x \leq 13.99$	$14 \leq x < 16.99$	$x \geq 17$

Since these methods use a lexicon of difficult words to make their predictions, and because such lexicon was not available to us, we manually inserted the excerpts into online platforms for calculating the readability.[3]

For the deep learning methods we selected the BERT transformer architecture due to its popularity and high performance on a wide range of natural language processing tasks [14].[4] Moreover, the model achieves state-of-the-art performance on both the WeeBit [5] and the OneStopEnglish [6] benchmarks.

[3] The platforms we used are: `goodcalculators.com/flesch-kincaid-calculator`, `charactercalculator.com/dale-chall-readability`, `gunning-fog-index.com`

[4] For an introduction to BERT we refer the reader to this blog post `https://jalammar.github.io/illustrated-bert/`.

We used a pre-trained `bert-base-cased` implementation provided by HuggingFace in their `transformers` library.[5] We fine-tuned the model on the OneStopEnglish dataset using an 80-20 split for training and testing. We ran experiments for 3, 5, 7, 10 and 15 epochs and recorded the test accuracy score. We report the test accuracy in Table 3.

Table 3. Test accuracy score of fine-tuned BERT model on OneStopEnglish test set

Num. epochs	Test accuracy
3	84%
5	92%
7	93%
10	95%
15	96%

We recorded the labels predicted by each of the methodologies above and calculated the accuracy on our scientific writing corpus. We report our findings in the next section.

3 Results

We report the accuracy of each of the four methodologies on our scientific writing annotated corpus in Table 4. The scores were calculated using the `accuracy score` from the `sklearn` library in Python.[6] As we can observe, all of the scores are rather low. Noticeably, BERT (7 epochs) and the Gunning Fog index produced similar results, and are the only methods that perform better than random guessing.[7]

Table 4. Accuracy scores on scientific writing test set. The best scores are in bold.

Method	Accuracy
Flesch Reading Ease score	0.31
Dale-Chall readability formula	0.22
Gunning Fog Index	**0.36**
BERT (3 epochs)	0.2
BERT (5 epochs)	**0.36**
BERT (7 epochs)	0.27
BERT (10 epochs)	0.31
BERT (15 epochs)	0.27

[5] `https://huggingface.co/bert-base-cased`
[6] Documentation details available at `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html`.
[7] Recall that random guessing for a three-class classification task is 33%.

We analyzed the errors for each of the methods. We present in Table 5 the confusion matrix of the Flesch Reading Ease score. Noticeably, the majority of the mistakes for this method are due to overestimating the difficulty level of the text. For example a text with gold label "medium" is estimated to be hard 8 times compared to a text with gold label "easy", for which this mistake occurs only once. This could be resolved by adjusting the conversion of the scales. However, this would not be a reasonable strategy, as it is based on the unrealistic assumption that all testing data we are interested in predicting will have gold labels with which the scale can be adjusted. Hence, we conclude that the Flesch Reading Ease score cannot provide an accurate predictor for scientific writing.

Table 5. Confusion matrix of the Flesch Reading Ease

True\Flesch	easy	intermediate	hard	Total
easy	5	8	9	22
intermediate	1	4	8	13
hard	0	5	5	10
Total	6	17	22	45

Table 6 reports the confusion matrix of the Dale-Chall readability formula. Note that the performance of this method is very low in almost all cases and the formula fails to provide with a good predictor for scientific writing.

Table 6. Confusion matrix of the Dale-Chall readability formula

True\Dale	easy	intermediate	hard	Total
easy	4	9	9	22
intermediate	6	4	3	13
hard	5	3	2	10
Total	15	16	14	45

The Gunning Fog Index (Table 7) achieves 36% accuracy on our test set, and is thus the best performing along with BERT. Noticeably, this method makes substantially more errors when the true label of the text is "easy". In this case it classifies the majority of the texts as being "hard" instead, which is not very intuitive: we would expect the formula to confuse more similar values, instead of the extremes. If we could reduce this error, the formula would perform better on scientific texts.
BERT also achieves 36% accuracy on our test set when trained for 5 epochs. Table 8 illustrates the corresponding confusion matrix. Notice that it classifies over one half of the cases as hard, even though only a fourth of the texts are actually hard according to the gold labels. Because of its preference for the hard exam-

Table 7. Confusion matrix of the Gunning Fog Index

True\Gunning	easy	intermediate	hard	Total
easy	6	5	11	22
intermediate	3	5	5	13
hard	4	1	5	10
Total	13	11	21	45

ples it performs worse when classifying normal and easy texts correctly. This could potentially be improved with more regularization in order to discourage the model from preferring one label over the others.

Table 8. Confusion matrix of BERT trained for 5 epochs

True\Bert	easy	intermediate	hard	Total
easy	7	1	14	22
intermediate	3	3	7	13
hard	1	3	6	10
Total	11	7	27	45

Figure 2 illustrates the stark difference in performance for BERT when tested with the OneStopEnglish dataset and our scientific writing corpus test set. The evident lower accuracy signals poor generalization abilities of this model.

4 Conclusion

With respect to our research questions we can conclude that

1. The traditional readability formulas we used are currently unable to provide accurate predictions for scientific writing texts
2. The state-of-the-art deep learning models we tested show poor generalization skills for scientific writing texts

Moreover we can observe that our initial hypothesis was incorrect, since the Gunning Fog Index was able to match performance with BERT, and out of the five BERT models that we fine-tuned none was higher than the other formulas (see Table 4).

We believe that the poor accuracy of the readability formulas is mostly due to two reasons. Firstly, the lexicons for hard words that they utilize are unlikely to cover scientific jargon. Secondly, the linguistic features they base their calculations on are too shallow. The latter criticism of the formulas is supported by multiple studies that highlighted the limitations of the used features [3, 15]. As for the deep learning methods, we believe that better results could be achieved if they were fine-tuned on more diverse and larger datasets for ARA.

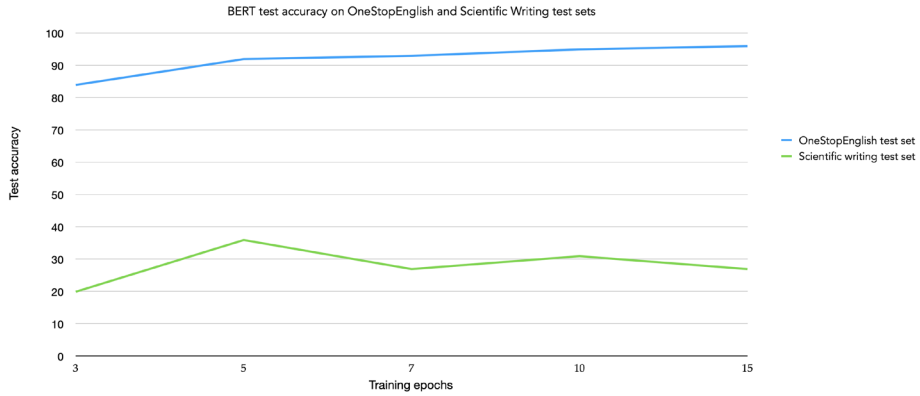


Fig. 2. Difference in test accuracy for BERT model on OneStopEnglish dataset and the scientific writing corpus, by number of training epochs. The distance between the two lines represents the lack of generalization abilities of the model.

In this paper we have analyzed different methods to assess readability. To conduct this analysis we created a test set for readability of scientific papers. We obtained the gold labels for this test set through human annotations collected in our survey. We compared the predictions from the Flesch Readability Ease, the Dale-Chall, the Gunning Fog indexes, and a BERT-based classifier to the gold labels. Based on the accuracies of each method, we concluded that the Gunning Fog Index and BERT perform the best.

However their accuracy is only slightly better than guessing. Therefore neither of these methods should be used to predict the readability of scientific papers. We recommend improving these formulas for scientific texts instead. For the analytical formulas further research is necessary to test different linguistic features and their performance on scientific articles. For the deep learning algorithms, more elaborate datasets should be developed. These datasets should contain enough humanly annotated samples from scientific texts.

References

1. Rebekah George Benjamin. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88, March 2012.
2. Donald P Hayes. The growing inaccessibility of science. *Nature*, 356(6372):739–740, April 1992.
3. Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359, 2017.

4. Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, April 2021.
5. Sowmya Vajjala and Detmar Meurers. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada, June 2012. Association for Computational Linguistics.
6. Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
7. Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. The readability of scientific texts is decreasing over time. *eLife*, 6:e27725, September 2017.
8. Adrian Barnett and Zoe Doubleday. The growth of acronyms in the scientific literature. *eLife*, 9:e60080, July 2020.
9. Lars Wissler, Mohammed Almashraee, Dagmar Monett, and Adrian Paschke. The gold standard in corpus annotation. June 2014.
10. J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
11. Rudolf F. Flesch. A new readability yardstick. *The Journal of applied psychology*, 32(3):221–33, June 1948.
12. Jeanne Sternlicht Chall and Edgar Dale. Readability revisited: The new dale-chall readability formula. 1995.
13. Bartosz Broda, Maciej Ogrodniczuk, Bartłomiej Nitoń, and Włodzimierz Gruszczyński. Measuring readability of polish texts: Baseline experiments. May 2014.
14. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
15. Bruce W. Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. Pushing on text readability assessment: A transformer meets handcrafted linguistic features, 2021.