Gender and Race in Carolina Digital Repository Content Methodology Review

Rebekah Kati
Institutional Repository Librarian
University of North Carolina at Chapel Hill

Introduction

In 2021, I conducted a review of subject area, gender and race representation in three Carolina Digital Repository (CDR) projects. The goal of the review was to determine if the content produced by the three projects were representative of the demographics of the University of North Carolina at Chapel Hill (UNC-CH). I used a publicly available, widely distributed list of faculty members who self-identified as Black, indigenous or a person of color and compared the list to CDR deposits. Also, I used a frequently cited API tool to determine author gender, based on recommendations from several bibliometrics studies.

At the end of the review, I recommended further research and reflection on ways to identify gender and race of CDR authors in an accurate and ethical manner. This report represents the first step in that reflection, and it will be an ongoing and iterative process. For this assessment, I first looked at studies which estimated race and/or gender composition of their subjects. I then categorized the methodologies to determine the most used. In this follow-up review, I will report on the results of my investigation into alternate methods to identify gender and race of authors, provide an evaluation of the previous study based on my findings and provide a recommendation for future work.

Gender

I looked at 23 studies in a variety of fields which estimated the gender of subjects based on their names. For each study, I locate the methodology section and noted the methods by which the study authors determined gender. The primary methodologies are grouped in the table below. Articles which use multiple methods have been grouped by the primary method.

Method	Number of Studies
API (e.g. gender API, genderize.io etc.)	15
Photo	2
Pronouns	5
Salutation	1

Most studies used an API tool to estimate gender. These tools rely on a large dataset of given names matched to gender. Datasets can be created from publicly available data sources, such as the United States Social Security Administration. The API will match the user input to the dataset and return a result of "male", "female" or "unknown" with a measure of confidence expressed as a percentage. Typically, the study authors determined a threshold percentage by which they assume the predictions to be accurate, which ranged from 70% to over 90%. Matches lower than the threshold percentage are either discarded or validated using another method.

Several studies noted the drawbacks of API-based tools:

1. They represent a binary view of gender and are therefore exclusionary to non-binary individuals.
2. The data sources are highly dependent on nationality. Data sources can lack coverage, particularly for East Asian names in Western datasets. Additionally, names can be associated with different genders depending on nationality (e.g., Andrea is a common female name in the United States, but a common male name in Italy.)
3. They do not predict gender neutral names accurately.

The second most popular method of gender assessment is preferred pronouns. In this method, researchers performed web searches for subjects' names to identify the pronouns that they used on social media and departmental websites. This method assumes that the subjects have a web or social media presence, which may not be true for subjects in historic datasets. Additionally, it assumes that the subject's publicly used pronouns align with their true gender identity.

Two studies used photographs as the primary method to assess gender. The researchers used publicly available websites such as department web pages to obtain photographs of the subjects and assessed their gender presentation. This method assumes that photographs are available for every subject and that the subject's public gender presentation reflects their true gender. Additionally, the method relies on the researcher's preconceptions of gender presentation, which could introduce bias.

Finally, one study relied on gendered salutations such as "Mr.", "Mrs.", "Miss" or "Ms." As with the previous two methods, this method assumes that the subject's publicly used salutation aligns with their true gender. This method also has limited utility in an academic environment where the gender-neutral salutations "Dr." or "Prof." are common.

Race

I looked at 15 studies in a variety of fields which assessed the racial makeup of subjects. For each study, I located the methodology section and noted the methods by which the study authors determined race. The primary methodologies are grouped in the table below. Articles which use multiple methods have been grouped by the primary method.

Method	Number of Studies
Self-identified	9
Photo	2
External dataset	2
Unknown	2

Most studies used self-identification to determine race, in which the subjects indicated their race based on pre-written categories on a questionnaire or similar. This method assumes that a questionnaire can be viably administered, and that the subject's racial identity falls neatly into the provided categories.

Two studies used photos to determine the race of their subjects. As with the gender assessment, this method assumes that photographs will be available for all subjects and relies on the assessor's preconceptions of race, which may be biased.

An additional two studies used external datasets which listed the subjects' race but did not identify how the external datasets assessed race to determine those categories. Finally, two studies indicated that race was assessed, but did not provide explicit information about their data gathering or assessment process.

Recommendations

Gender

An API-based approach was by far the most used method for assessing gender, which aligns with my methodology in the first round of assessment. However, the drawbacks noted in the section above are significant and need to be acknowledged when performing an assessment. Furthermore, when an API-based approach is used, it is best to set the threshold percentage high and to use an alternate approach to manage outliers. However, using this approach on a large dataset such as the CDR's dataset will generate a good deal of manual work that will need to be managed and accounted for.

An approach based on the subject's preferred pronouns is preferable to an API-based approach as it reflects the subject's gender identity. Given the large size and age of the dataset used for the CDR analysis, this method will eliminate many older articles since authors of older articles will be less likely to have a current web or social media presence. It may be useful to analyze a sample set to establish a baseline and to assess potential issues.

Race

Self-identification was by far the most utilized method to assess race. This aligns with the methodology of the previous assessment, since the subjects were provided the option to self-identify as a member of a racial minority group. Since the CDR does not require users to provide demographic information upon deposit and we have no plans to do so, the previous methodology is the best way to determine race of depositors. However, the list only includes faculty who chose to self-identify as a member of a minority group and does not reflect a comprehensive view of the university. Due to de-identification, it is unlikely that an external dataset from university HR or research office would provide enough information for further assessment. Assessment based on photographs would introduce bias and would be ethically dubious at best.

Therefore, it may be advisable to shift focus from spotlighting work authored by BIPOC authors to a topic-based approach. In this approach, I would identify keywords and research topics relating to minority populations and verify that eligible research was available in the CDR. One such approach was piloted at Virginia Tech, in which the researchers compiled a controlled vocabulary describing minority groups and used the terms to search the university's website and institutional repository.[1] These searches provided a basis of comparison to determine if research on a particular subject was being performed at the university and if so, whether its outputs were deposited in the institutional repository.

We hope that the approaches above will be a first step toward broadening the subject area, race, and gender focus of the CDR, which will bring the CDR more in line with UNC Libraries' Reckoning Initiative. We will continue to assess our progress and publicly publish updates on a yearly basis, as we have done with the Content Liberation projects and the CDR platform updates.

Studies Consulted

Gender

Abramo G, D'Angelo CA, Mele I. (2022) Impact of Covid-19 on research output by gender across countries. *Scientometrics*. 2022 Jan 27:1-16. https://dx.doi.org/10.1007/s11192-021-04245-x.

Bell ML and Fong KC. (2021). Gender Differences in First and Corresponding Authorship in Public Health Research Submissions During the COVID-19 Pandemic. *American Journal of Public Health*. 111, 159_163, https://doi.org/10.2105/AJPH.2020.305975

Chan, HF and Torgler, B. (2020). Gender differences in performance of top cited scientists by field and country. *Scientometrics* 125(3): 2421–2447. https://doi.org/10.1007/s11192-020-03733-w

Chary S, Amrein K, Soeteman DI, Mehta S, Christopher KB. (2021). Gender disparity in critical care publications: a novel Female First Author Index. *Annals of Intensive Care*. 11(1):103. https://dx.doi.org/10.1186/s13613-021-00889-3.

Gayet-Ageron A, Ben Messaoud K, Richards M, Schroter S. (2021). Female authorship of covid-19 research in manuscripts submitted to 11 biomedical journals: cross sectional study. *BMJ* 375 :n2288. https://dx.doi.org/10.1136/bmj.n2288

Heath JK, Clancy CB, Carillo-Perez A, Dine CJ. (2020). Assessment of Gender-based Qualitative Differences within Trainee Evaluations of Faculty. *Annals of the American Thoracic Society*. 17(5):621-626. https://dx.doi.org/10.1513/AnnalsATS.201906-479OC.

Hornstein P, Tuyishime H, Mutebi M, Lasebikan N, Rubagumya F, Fadelu T. (2022). Authorship Equity and Gender Representation in Global Oncology Publications. *JCO Global Oncology*. 8:e2100369. https://dx.doi.org/10.1200/GO.21.00369.

Ibrahim H, Abdel-Razig S, Stadler DJ, Cofrancesco J Jr, Archuleta S. (2019). Assessment of Gender Equity Among Invited Speakers and Award Recipients at US Annual Medical Education Conferences. *JAMA Network Open*. 2(11):e1916222. https://dx.doi.org/10.1001/jamanetworkopen.2019.16222.

[1] McMillan, Gail. (2021). Is the IR Storage or Showcase? Presentation for ACRL. http://hdl.handle.net/10919/103226

Lee SF, Redondo Sánchez D, Sánchez MJ, Gelaye B, Chiang CL, Wong IOL, Cheung DST, Luque Fernandez MA. (2021). Trends in gender of authors of original research in oncology among major medical journals: a retrospective bibliometric study. *BMJ Open*. 11(10):e046618. https://dx.doi.org/10.1136/bmjopen-2020-046618.

Maggio LA, Costello JA, Ninkov A, Frank JR, Artino Jr. AR (2022). The Voices of Medical Education Science: Describing the Published Landscape. *BioRxiv*. https://doi.org/10.1101/2022.02.10.479930

Malkinson TS, Terhune DB, Kollamkulam M, Guerreiro MJ, Bassett DS, Makin TR (2021). Gender Imbalance in the Editorial Activities of a Researcher-led Journal. *BioRxiv.* https://doi.org/10.1101/2021.11.09.467796

Mamtani M, Shofer F, Mudan A, et al. (2020). Quantifying gender disparity in physician authorship among commentary articles in three high-impact medical journals: an observational study. *BMJ Open*. 10:e034056. https://dx.doi.org/10.1136/bmjopen-2019-034056

Oliveira-Ciabati L, Santos LL, Hsiou AS, Sasso AM, Castro M, Souza JP. (2021). Scientific sexism: the gender bias in the scientific production of the Universidade de São Paulo. *Revista de Saude Publica*. 55:46. https://dx.doi.org/10.11606/s1518-8787.2021055002939.

Paul-Hus A, Mongeon P, Sainte-Marie M, Larivière V. (2020). Who are the acknowledgees? An analysis of gender and academic status. *Quantitative Science Studies.* 1 (2): 582–598. https://dx.doi.org/10.1162/qss_a_00036

Salter-Volz AE, Oyasu A, Yeh C, Muhammad LN, Woitowich NC. (2021). Sex and Gender Bias in Covid-19 Clinical Case Reports. *Frontiers in Global Women's Health*. 2:774033. https://dx.doi.org/10.3389/fgwh.2021.774033.

Squazzoni F, Bravo G, Farjam M, Marusic A, Mehmani B, Willis M, Birukou A, Dondio P, Grimaldo F. (2021). Peer review and gender bias: A study on 145 scholarly journals. *Science Advances*. 7(2):eabd0299. https://dx.doi.org/10.1126/sciadv.abd0299.

Thomas EG, Jayabalasingham B, Collins T, Geertzen J, Bui C, Dominici F. (2019). Gender Disparities in Invited Commentary Authorship in 2459 Medical Journals. *JAMA Network Open*. 2(10):e1913682. https://dx.doi.org/10.1001/jamanetworkopen.2019.13682.

Chen T-HK, Seto KC (2022). Gender and authorship patterns in urban land science. *Journal of Land Use Science*. https://dx.doi.org/10.1080/1747423X.2021.2018515

Wang LL, Stanovsky G., Weihs L., Etzioni O. (2021). Gender Trends in Computer Science Authorship. *Communications of the ACM*. 64(3), 78-84. https://dx.doi.org/10.1145/3430803

Wehner MR, Li Y, Nead KT. (2020). Comparison of the Proportions of Female and Male Corresponding Authors in Preprint Research Repositories Before and During the COVID-19 Pandemic. *JAMA Network Open*. 3(9):e2020335. https://dx.doi.org/10.1001/jamanetworkopen.2020.20335.

Wright KM, Edberg D, Wheat S, Clements DS. (2019). Prevalence of Women Authors in Family Medicine Literature. *JAMA Network Open*. 2(11):e1916029. https://dx.doi.org/10.1001/jamanetworkopen.2019.16029.

Yamamoto J, Frachtenberg E. (2022). Gender Differences in Collaboration Patterns in Computer Science. *Publications*. 10(1):10. https://doi.org/10.3390/publications10010010

Zhang, L., Shang, Y., Huang, Y. et al. (2022). Gender differences among active reviewers: an investigation based on publons. *Scientometrics* 127, 145–179. https://dx.doi.org/10.1007/s11192-021-04209-1

Race

Ashford MT, Eichenbaum J, Williams T, Camacho MR, Fockler J, Ulbricht A, Flenniken D, Truran D, Mackin RS, Weiner MW, Nosheny RL. (2020). Effects of sex, race, ethnicity, and education on online aging research participation. *Alzheimer's and Dementia Translational Research and Clinical Interventions*. 6(1):e12028. https://dx.doi.org/10.1002/trc2.12028.

Bauer H, Gebresenbet F, Kiki M, Simpson L, Sillero-Zubiri C (2019) Race and Gender Bias in the Research Community on African Lions. Frontiers in Ecology and Evolution 7:24. https://dx.doi.org/10.3389/fevo.2019.00024

Dupree CH, Torrez B, Obioha O, Fiske ST. (2021). Race-status associations: Distinct effects of three novel measures among White and Black perceivers. Journal of Personality and Social Psychology. 120(3):601-625. https://dx.doi.org/10.1037/pspa0000257.

Evans KE, Munson B, Edwards J. (2018). Does Speaker Race Affect the Assessment of Children's Speech Accuracy? A Comparison of Speech-Language Pathologists and Clinically Untrained Listeners. *Language, Speech, and Hearing Services in Schools*. 49(4):906-921. https://dx.doi.org/10.1044/2018_LSHSS-17-0120.

Fort, R., & Gill, A. (2000). Race and Ethnicity Assessment in Baseball Card Markets. *Journal of Sports Economics*, 1(1), 21–38. https://doi.org/10.1177/152700250000100103

Freeman A. (2018). Mathematics Self-Efficacy and the Smarter Balanced Assessment: An Intersection of Race, Socioeconomic Status and Gender. [Doctoral dissertation, George Fox University]. https://digitalcommons.georgefox.edu/edd/120

Ginther DK, Basner J, Jensen U, Schnell J, Kington R, Schaffer WT. (2018). Publications as predictors of racial and ethnic differences in NIH research awards. *PLoS One*. 13(11):e0205929. https://dx.doi.org/10.1371/journal.pone.0205929.

Jawitz J. (2012) Race and assessment practice in South Africa: understanding black academic experience, *Race, Ethnicity and Education.* 15:4, 545-559, https://dx.doi.org/10.1080/13613324.2011.645568

Kozlowski, D, Lariviere V, Sugimoto CR, Monroe-White T. (2022). Intersectional inequalities in Science. *Proceedings of the National Academy of Sciences.* 119(2):e2113067119. https://dx.doi.org/10.1073/pnas.2113067119

Lisnic, R., Zajicek, A., & Morimoto, S. (2019). Gender and Race Differences in Faculty Assessment of Tenure Clarity: The Influence of Departmental Relationships and Practices. *Sociology of Race and Ethnicity*, 5(2), 244–260. https://doi.org/10.1177/2332649218756137

Pardoel K. (2020). An Examination of the Influence of Gender and Race on Dynamic Risk Assessment. [Doctoral dissertation, Carleton University]. https://curve.carleton.ca/55223b8f-a06a-4765-b591-95d6d9a639b9

Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, Hripcsak G, Smith S, Tatonetti NP, Vawdrey DK. (2019). Challenges with quality of race and ethnicity data in observational databases. *Journal of the American Medical Informatics Association*. 26(8-9):730-736. https://dx.doi.org/10.1093/jamia/ocz113.

Punti G, Dingel M. (2021). Rethinking Race, Ethnicity, and the Assessment of Intercultural Competence in Higher Education. *Education Sciences*. 11(3):110. https://doi.org/10.3390/educsci11030110

Silva MND, Monteiro JCDS. (2020). Self-esteem assessment of young female university students according to race/skin color criteria. *Revista Latino-Americana de Enfermagem*. 28:e3362. https://dx.doi.org/10.1590/1518-8345.3866.3362.

Willis, DE, Andersen, JA, Bryant-Moore, K, et al. (2021). COVID-19 vaccine hesitancy: Race/ethnicity, trust, and fear. *Clinical and Translational Science*. 14: 2200– 2207. https://dx.doi.org/10.1111/cts.13077