Journal Requirements:

1. Please ensure that your manuscript meets PLOS ONE's style requirements, including those for file naming. The PLOS ONE style templates can be found at

https://journals.plos.org/plosone/s/file?id=wjVg/PLOSOne_formatting_sample_main_body.pdf and

https://journals.plos.org/plosone/s/file?id=ba62/PLOSOne_formatting_sample_title_authors_affiliations.pdf

Answer:

We have modified the format to adapt it to PLOS ONE style.

2. Thank you for stating the following financial disclosure:

"VL acknowledges funding from the Canada Research Chairs program, https://www.chairs-chaires.gc.ca/, (grant # 950-231768), DK acknowledges funding from the Luxembourg National Research Fund, https://www.fnr.lu/, under the PRIDE program (PRIDE17/12252781)."

Please state what role the funders took in the study. If the funders had no role, please state: "The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript."

If this statement is not correct you must amend it as needed.

Please include this amended Role of Funder statement in your cover letter; we will change the online submission form on your behalf.

Answer:

We have add the amendment in the cover letter.

Reviewer #1:

I wonder about the section regarding imputation of missing data. Maybe, the authors could provide advice on whether imputation should be done at all. It seems to be taken for granted that some imputation method should be used. Is imputation from a proper distribution better than no imputation at all? Authors without proper assignment could be just removed from the data set. What would be the advantages and disadvantages?

Answer:

We thank the reviewer for this valuable question. Indeed, this was taken for granted in the original submission, but constitutes a very important debate. Both the imputation and the omission of unknown cases can generate a bias if the distribution of unknown cases is not random. We addressed this issue in the (new) Fig 6, and this new paragraph:

> Nevertheless, this type of imputation can also introduce new biases. If the missing family names correlate with a specific racial group, then the known cases cannot be considered a random sample of the data, and their mean will be biased toward those groups that have fewer unknown names. Knowing which group has more unknown cases is in principle an impossible task. Nevertheless, it is possible to infer this, considering the citizenship status of authors. Authors that are temporary visa holders in US are more likely to have a family name that doesn't appear on the census. The

Survey of Earned Doctorates provides information on doctorate recipients, by ethnicity, race, and citizenship status between 2010 and 2019 [38]. Fig 6 shows the average proportion of Temporary Visa Holders among Earned Doctorates from each racial group. This can be seen as a proxy of the distribution of authors by race and citizenship status. There is a large majority of Asian authors that are migrants, followed by a 30% of Hispanic authors, 19% of Black authors and 11% of White authors. Imputing by the mean of the known authors would also underestimate Asian authors, and partially too Hispanic authors, while overestimating White authors. Nevertheless, omitting the missing cases would have the same effect on the overall distribution, given that the imputation by the mean does not change the aggregate proportion of each group. There is no perfect solution for this, as the distribution shown on Fig 6 is only a proxy of the problem. Therefore, it is important to acknowledge this potential bias on the result, both if the imputation is used or if the missing cases are omitted.

Reviewer #2:

152-158: At this point, it was difficult for me to get an idea of what the simulated data is used for. A sentence after "First, to test the interaction between given and family names distributions, we simulate a dataset that covers most of the possible combinations" could help to clarify this (e.g. something like "This step is only used to determine how to combine given and family names for inferring race").

Answer:

We added a clarifying sentence (170-172 in the highlighted version)

169-170: "Questions now include both racial and ethnic origin, placing "Hispanic" outside racial categories. The racial categories in both datasets include Hispanic as a category, ...". At first sight, this sounds contradictory ("Hispanic" is not in racial categories, and at the same time "Hispanic" is used as a category). I would suggest to clarify here that "Hispanic" is not a racial category in the original US Census data, but you use it as a racial category in your datasets. There are also no quotes around "Hispanic" in line 170, while this is usually the case in line 169.

Answer:

We thank the reviewer for the suggestion, the data sources used on this manuscript add 'hispanic' as a racial category. We added a clarifying sentence (184-188 in the highlighted version)

186-187: This implies that only first authors are considered for your analyses. This restriction should be mentioned explicitly, and also why you chose to do so (and did not include other author positions).

Answer:

Indeed, we only used first authors to be sure they were US-based. We added a clarification in lines 201-205 of the highlighted version

193-222: This part seems to better fit in the "Methods" section than the "Data" section. Unless there are good reasons to keep it in the "Data" section, you may want move this part.

Answer:

We moved this part to the Methods section.

226: In the "Methods" section, it is unclear which particular approaches you finally use in your empirical analyses. In particular, which of the three weighting schemes did you finally use for your analyses? I think a concise list of the approaches you use would be helpful for the reader.

Answer:

We added a list of the methods used on the experiments at the end of the Methods section (317-323 of the highlighted version)

252: Use "n" instead of "c" in the summation notation for consistency with the formula in line 245 (or vice versa).

Answer:

We thank the reviewer for noticing this, we fixed the inconsistency

254: "for both given (family) names" looks like it is a measure for two given names or two family names. But as far as I understand it, this weight combines the given and the family name of one person. Should this be "for both given and family names"?

Answer:

We fix the sentence to make it more clear

259: Which value did you finally use for exp?

264-272: Which color pattern should be observed in order to have a good approach? Have you tried exponent values > 2? If not, why?

Answer:

We thank the reviewer for this comment, we explored values of the exponent 1 and 2. We further clarify why in the sentence 282-286 of the highlighted version.

290: ";" -> "."

Answer:

We fixed the typo

304-307: A more detailed explanation of how the given name distribution has been normalized and how this affects the results would be helpful here.

Answer:

We agree with the reviewer that a more detailed explanation was needed. We add a clarification in lines 337-346 of the highlighted version

Figure 3: How are the values on the y-axis (ratio) calculated, and how does this measure over-/underrepresentation? The frequencies of given/family names that are shown in the upper plots provide important information, but the distribution is difficult to inspect for thresholds > 90%. Can this be visualized in a form that better shows this part of the distribution (e.g. in an extra figure)?

Answer:

We add further explanation of this on lines 353-356 of the highlighted version. We have also divided Figure 3 into A and B, where A is the original figure, and B shows a detailed version for thresholds above 90%

399-416: It is a good point and convincingly shown by the results presented here that simply imputing based on the Census data should be avoided. But imputing based on the distributions in the bibliometric data for known names may also introduce biases. This would be the case if the probability for missing names correlates with the race category (a reason for this might be the development over time of both the probability to have full names in the database and the distribution across race categories). I would argue one has to be very cautious when imputing bibliometric data, because usually these data only provide a limited amount of metadata that can be used for imputation. Given the usually large number of cases in bibliometric data, it is probably not necessary to impute data in most cases. I would argue that it is more important and transparent to discuss possible biases for a particular research question introduced due to missing data than trying to impute the data. I think your results also provide a very good basis for such a discussion with regard to inferring race.

Answer:

We thank the reviewer for raising this very important issue. As we mentioned for reviewer #1, we agree this was not properly address and it can be a potential source of bias. Both the imputation by the mean and the omission of missing cases can be introduced bias if the distribution of missing name is not random. We added an explanation for this in lines 461-478 of the highlighted version, and we also added Fig 6. That shows the potential non-randomness of missing cases given by the citizenship status of authors.