Manuscript Title

This manuscript (<u>permalink</u>) was automatically generated from <u>greenelab/disparities opinion@5f7013e</u> on November 3, 2021.

Authors

- John Doe

Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- Jane Roe
 - D XXXX-XXXX-XXXX · janeroe

Department of Something, University of Whatever; Department of Whatever, University of Something

Recognition by peers or the general public can greatly affect the trajectory of a scientist's career since each form of recognition paves the way to future accolades. Regardless of an individual's merit, biases can skew which scientists are considered for recognition. Audits can combat this by identifying neglected groups and hold the recognizing body accountable.

Auditing efforts have shown that they can help reduce disparities. Journalists have found performing self-audits to quantify gender-based representation in their sources can help them approach parity. [1/,2/].

In award recognition, an audit of the International Society for Computational Biology (ISCB) honorees revealed significant under-representation of people with East Asian name origins and over-representation of US-affiliated scientists. After this study was publicized, ISCB's next honorees had the highest mean predicted probability of having an East Asian name of any previous year. Additionally, the nominating committee inducted the first China-based fellow [3].

Audits can also provide evidence for specific actions. A recent study of disparities in scientific journalism found that predicted East Asian names were under-quoted and under-mentioned [4]. However, this disparity was almost completely removed when considering quoted or mentioned people with a US-affiliated paper cited in the article. This suggests a regional bias, arguing for more regionally focused or co-located journalists.

Audits of representation would ideally be done with all individuals self-reporting gender, ethnicity, and other identifications [5]. While this may be possible prospectively, surveying for self-identification is impractical for large groups and often impossible retrospectively [6/]. Computationally-derived predictions allow for audits on a scale that would not be possible otherwise. Numerous tools exist to algorithmically infer gender, nationality, and ethnicity information using only the feature most likely to be non-missing in a relevant dataset: an individual's name. [7,8/] Most of these models are highly scalable, allowing auditors to define the scope of their target group and background population as broadly as needed.

Prediction models are not a panacea; several factors limit both their accuracy at recapitulating self-reported information and their ability to address the underlying motivating questions of diversity audits. For instance, gender associations of a given name can vary by culture, potentially biasing gender predictions where additional information is not available [9]. Also, most gender prediction models are trained on binary gender labels, which occludes assessing the representation of transgender, non-binary, and intersex individuals [10].

Proxy predictions of ethnicity via name origin are more difficult still; choosing categories to probabilistically predict on is non-trivial and difficult to define. Furthermore, there is no one-to-one mapping between having a name from a linguistic group and belonging to a minoritized or underrepresented group. Colonialism, immigration, and structural racism have affected most groups' linguistic history and inclusion or exclusion from scientific communities in complex ways that are nearly impossible to parse from names alone. For instance, classifiers are usually unable to distinguish if names of Hispanic origin come from the Iberian Peninsula or from Latin America [3]. If a target and background population had identical probabilistic proportions of Hispanic names, but the target group primarily consisted of Iberian individuals, underrepresentation of Latin American scientists would go unnoticed in an exclusively computational analysis.

Recognizing the aforementioned shortcomings, we propose the following recommendations for the creation and deployment of automated auditing tools: 1. *Transparency*. Publicly provide all tools, code, and data used in the analysis. This enables public scrutiny to those being audited and those whose data you are using. For confidential data, we recommend providing de-identified or aggregated data.

- 2. *Individuals know best.* Self-identified demographic information should be used in preference to algorithmic predictions.
- 3. *Aggregates only.* Audit results should not affect individuals. Analyses must focus on aggregate estimations and any intermediary individual predictions should not be used externally. In addition, analysts must be mindful of hidden subpopulations that may be obscured during aggregation.
- 4. *Inform the public.* While internal audits can help institutions reflect, the public must be given an opportunity to hold them accountable.

This manuscript is a template (aka "rootstock") for <u>Manubot</u>, a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (.md files in the /content directory) to the output you see below.

Basic formatting

| Bold text | | |
|----------------|---------------|--------------------|
| Semi-bold text | | |
| | Centered text | |
| | | Right-aligned text |

Italic text

Combined italics and bold

Strikethrough

- 1. Ordered list item
- 2. Ordered list item
 - a. Sub-item

- b. Sub-item
 - i. Sub-sub-item
- 3. Ordered list item
 - a. Sub-item
- List item
- List item
- · List item

subscript: H₂O is a liquid

superscript: 2¹⁰ is 1024.

unicode superscripts⁰¹²³⁴⁵⁶⁷⁸⁹

unicode subscripts₀₁₂₃₄₅₆₇₈₉

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to <u>editing</u> and <u>version</u> <u>control</u>.

Line break without starting a new paragraph by putting two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

Heading 6



Horizontal rule:

Heading 1's are recommended to be reserved for the title of the manuscript.

Heading 2's are recommended for broad sections such as Abstract, Methods, Conclusion, etc.

Heading 3's and Heading 4's are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

<u>Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah</u>

Link with text

Link with hover text

Link by reference

Citations

Citation by DOI [11].

Citation by PubMed Central ID [12].

Citation by PubMed ID [13].

Citation by Wikidata ID [14].

Citation by ISBN [15].

Citation by URL [16].

Citation by alias [17].

Multiple citations can be put inside the same set of brackets [11,15,17]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [12,13,17,18].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure 1

Figure 2

```
Figure 3

Figure 4

Table 1

Equation 1

Equation 2
```

Quotes and code

Quoted text

Quoted block of text

Two roads diverged in a wood, and I—I took the one less traveled by, And that has made all the difference.

Code in the middle of normal text, aka inline code.

Code block with Python syntax highlighting:

```
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

Figures



Figure 1: A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.



Figure 2: An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.



Figure 3: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.



Figure 4: A vector .svg image loaded from GitHub. The parameter sanitize=true is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image.

Tables

Table 1: A table with a top caption and specified relative column widths.

| Bowling Scores | Jane | John | Alice | Bob |
|----------------|------|------|-------|-----|
| Game 1 | 150 | 187 | 210 | 105 |
| Game 2 | 98 | 202 | 197 | 102 |
| Game 3 | 123 | 180 | 238 | 134 |

Table 2: A table too wide to fit within page.

| | Digits 1-33 | Digits 34-66 | Digits 67-99 | Ref. |
|---|---------------------------------|--------------|--------------|-------------|
| р | 3.141592653589 8462643383279 | | | I niday org |
| е | 2.718281828459 5360287471352 | | | nasa gov |

 Table 3: A table with merged cells using the attributes plugin.

| | Colors | |
|-------|------------|------------------|
| Size | Text Color | Background Color |
| big | blue | orange |
| small | black | white |

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \tag{1}$$

An equation too long to fit within page:

$$x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$$
(2)

Special

▲ WARNING The following features are only supported and intended for .html and .pdf exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as .docx .

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot. Manubot Manubot. Manubot. Manubot. Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot attributes plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

Manubot Manubot.

Available background colors for text, images, code, banners, etc:

white lightgrey grey darkgrey black lightred lightyellow lightgreen lightblue lightpurple red orange yellow green blue purple

Using the Font Awesome icon set:



Light Grey Banner
useful for general information - manubot.org

1 Blue Banner

useful for important information - manubot.org

♦ Light Red Banner useful for *warnings* - <u>manubot.org</u>

References

1. I Analyzed a Year of My Reporting for Gender Bias (Again)

Adrienne LaFrance

The Atlantic (2016-02-17) https://www.theatlantic.com/technology/archive/2016/02/gender-diversity-journalism/463023/

2. I Spent Two Years Trying to Fix the Gender Imbalance in My Stories

Ed Yong

The Atlantic (2018-02-06) https://www.theatlantic.com/science/archive/2018/02/i-spent-two-years-trying-to-fix-the-gender-imbalance-in-my-stories/552404/

3. Analysis of scientific society honors reveals disparities

Trang T Le, Daniel S Himmelstein, Ariel A Hippen, Matthew R Gazzara, Casey S Greene *Cell Systems* (2021-09) https://doi.org/gmhq49

DOI: <u>10.1016/j.cels.2021.07.007</u> · PMID: <u>34555325</u>

4. Analysis of scientific journalism in <i>Nature</i> reveals gender and regional disparities in coverage

Natalie R Davidson, Casey S Greene

Cold Spring Harbor Laboratory (2021-06-22) https://doi.org/gkscd5

DOI: 10.1101/2021.06.21.449261

5. Ten Simple Rules to Achieve Conference Speaker Gender Balance

Jennifer L Martin

PLoS Computational Biology (2014-11-20) https://doi.org/gf853n

DOI: 10.1371/journal.pcbi.1003903 · PMID: 25411977 · PMCID: PMC4238945

6. How to Ethically and Responsibly Identify Gender in Large Datasets

MediaShift

(2014-11-21) http://mediashift.org/2014/11/how-to-ethically-and-responsibly-identify-gender-in-large-datasets/

7. Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records

Kosuke Imai, Kabir Khanna

Political Analysis (2017-01-04) https://doi.org/f8ntmv

DOI: 10.1093/pan/mpw001

8. **Genderize.io | Determine the gender of a name** https://genderize.io/

9. A Data-driven Approach to Studying Given Names and their Gender and Ethnicity Associations

Shervin Malmasi

Proceedings of the Australasian Language Technology Association Workshop 2014 (2014-11) https://aclanthology.org/U14-1021

10. Racial and ethnic imbalance in neuroscience reference lists and intersections with gender

Maxwell A Bertolero, Jordan D Dworkin, Sophia U David, Claudia López Lloreda, Pragya Srivastava, Jennifer Stiso, Dale Zhou, Kafui Dzirasa, Damien A Fair, Antonia N Kaczkurkin, ... Danielle S Bassett

Cold Spring Harbor Laboratory (2020-10-12) https://doi.org/gj7mdc

DOI: <u>10.1101/2020.10.12.336230</u>

11. Sci-Hub provides access to nearly all scholarly literature

Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene

eLife (2018-03-01) https://doi.org/ckcj

DOI: <u>10.7554/elife.32822</u> · PMID: <u>29424689</u> · PMCID: <u>PMC5832410</u>

12. Reproducibility of computational workflows is automated using continuous analysis

Brett K Beaulieu-Jones, Casey S Greene

Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMCID: 10.1038/nbt.3780/ · PMCID: 10.1038/nbt.3780/ · PMCID:

13. **Bitcoin for the biological literature.**

Douglas Heaven

Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888

DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

14. Plan S: Accelerating the transition to full and immediate Open Access to scientific publications

cOAlition S

(2018-09-04) https://www.wikidata.org/wiki/Q56458321

15. **Open access**

Peter Suber

MIT Press (2012)

ISBN: 9780262517638

16. Open collaborative writing with Manubot

Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter

Manubot (2020-05-25) https://greenelab.github.io/meta-review/

17. Opportunities and obstacles for deep learning in biology and medicine

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, ... Casey S Greene

Journal of The Royal Society Interface (2018-04-04) https://doi.org/gddkhn DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

18. **Open collaborative writing with Manubot**

Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter

PLOS Computational Biology (2019-06-24) https://doi.org/c7np

DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653