Social Divisions in Data

Simon Couch, 16 December 2019

1 Introduction

In her 1999 book "Moving Beyond Gender: Intersectionality and Scientific Knowledge," Patricia Hill Collins characterizes "positivist science," a paradigm under which scientific knowledge is produced that articulates specific research practices that lead to unbiased and objective truth. "Claiming the existence of absolute truths and an objective reality structured by invariant rules, positivist science argues that the underlying structure of social as well as physical phenomena can be uncovered." However, a well-established and growing body of feminist and intersectional work has crticized these methodologies for "using quantitative data in simplistic and superficial ways [and] improperly interpreting and overgeneralizing scientific findings" [2]. In this paper, I will characterize how this simplistic and superficial use of data has persisted in the positivist scientific endeavor by focusing in on the process of constructing datasets. Rather than viewing datasets as unargumentative, sterile, or "raw," I argue that datasets, and operations carried out on them, express the social conceptualizations held by their creators. Especially in reference to social categorizations and identities, naming columns; writing descriptions of these columns in codebooks; constructing categories within which subjects must identify (or be assigned); binning entries on a continuous scale into discrete categories; naming these new categories; constructing and naming new categories within columns from pre-existing categories; constructing and naming new columns from pre-existing columns; and assigning numerical (and by extension, ordinal) values to categories is a process of argumentation, claims-making, and boundary work. More specifically, in this paper, I examine how sex, gender, race, and ethnicity are named and encoded as variables in data, and how this process is patterned by the identities held by the datasets' creators.

2 Literature Review

2.1 Sex and Gender

The distinction between sex and gender, acknowledging the presence of feminized and masculinized socialization accompanying a more purely biological "sex," arose from feminist theory in the 1970s. In reference to Gayle Rubin, Richardson writes "Rubin distinguished between the biological category of 'sex' (typically, male or female) and the social roles and expectations of 'gender' (such as heterosexual masculinity and femininity). The sex/gender distinction analytically separates the anatomy and physiology of males and females (sex) from the behavioral and cultural expectations associated with the ideals of masculinity and femininity (gender)" [11]. This definition is not historical, too. As Springer et al. write in 2012, "The IOM, drawing especially on definitions advanced by the World Health Organization (WHO) and the style manual of the Journal of the American Medical Association (JAMA), defines sex as 'The classification of living things, generally as male or female according to their reproductive organs and functions assigned by chromosomal complement' and gender as 'A person's self-representation as male or female, or how that person is responded to by social institutions based on the individual's gender presentation.'" Rather than encourage the reconciliation of the ways that these two interact, though, this distinction has cemented 'sex' within the jurisdiction of biologists, and isolated the biological and social into easily separable categories in which outward-flowing causality is granted only to sex. As the same authors write, "These definitions lend a superficial sense that sex and gender are distinct domains, even as they give causal and temporal priority to biology ('gender is rooted in biology' but sex is presumably pristine and emerges regardless of environment and experience)" [14]. While those utilizing gender and gendered socialization as variables in research must reconcile the effects of sex on gender, natural scientists are free to regard sex as a pure, self-evident, and unalterable truth. Richardson reflects on this change similarly, writing "The sex/gender distinction served to harden the notion of X and Y as 'sex itself'... The X and Y came to represent the necessary alter ego of gender fluidity, signifying what

nature intended the sexual fate of the infant to be" [11].

This is not to say, though, that considerable scholarship has not been devoted to contesting this supposed purity of sex as a biological category. As early as 1977, when feminist biologists Ethel Tobach and Betty Rosoff organized the inaugural conference "Genes and Gender" to take on recent pop science arguing the inevitability of 'sexed' social roles, scientists have critiqued the usefulness of this distinction [15]. Further, many scientists have carried out work within the confines of the positivist framework showing the biological effects of gendered socialization on physiology. In 2005, Fausto-Sterling showed these effects in regard to bone development [3]; Jordan-Young and Rumiati made a similar argument about the brain in 2012 [7]; many studies in neuroendicrinology demonstrate the biological effects of social status and identity [1, 4, 16]. Richardson writes that studies like these show that "gendered life experiences have material effects on the body. These effects show up, in turn, as biologically based 'sex differences'" [11]. In response to these critiques, scholars suggest that scientists "conceptualize sex/gender as a domain of complex phenomena that are simultaneously biological and social, rather than a domain in which the social and biological 'overlap'" [14].

Despite these critiques, though, many scientists continue to capitalize on this distinction, and subsequent prioritization of sex as a causal mechanism for differences, in research. In one example, four scientists presented research on brain research in a 2014 panel at Barnard College. In a talk early on in the session, Rae Silver says "when I talk about sex differences... I'm thinking of sex as a biological construct, and I'm thinking of the genetic, hormonal, and metabolic factors. I'm not at all thinking about gender role, or what society tells us is male and female typical, and I'm not at all thinking about gender identity—how we express our experiences of our sexuality—because none of this research speaks to that" [13]. Not only does Silver argue that sex effects can be isolated in her research, but that gender identity is literally defined as the expression of sex differences. I do not cite this example for its strikingness or severity, but precisely the opposite; the way that Silver presents this

research is archetypal of the positivist scientific endeavor's treatment of sex/gender. Indeed, in 2018, Hanvivsky et al. write that, in funding and publishing and funding guidelines for epidemiology research, "criteria fail to recognize the complexity of sex/gender, including the intersection of sex/gender with other key factors that shape health." Sex/gender are only sometimes mentioned, and when they are, "there is wide variation in how sex/gender are conceptualized and how researchers are asked to address the inclusion/exclusion of sex/gender in research." Principally, these funding agencies often emphasize representation in those carrying out scientific research while leaving the methodology used to conceptualize sex/gender effects in resesarch results unscrutinized. "[R]equirements that have been institutionalized within funding agencies tend to prioritize greater male/female equality in research teams and funding outcomes over considerations of sex/gender in research content and knowledge production" [5]. As has long been argued in feminist scholarship, representation of those holding marginalized identities in science research is absolutely an important endeavor. This need not come at the cost of criticality about the ways that these social identities are conceptualized in resulting knowledge production, though.

2.2 Race and Ethnicity

The analogous story for race/ethnicity is similarly just as much one of semantic inconsistency as it is of claims-making of biological jurisdiction. Initially, the instability across time and place of racial categories make the uselessness of race as a biological mechanism immediate. "Since its invention to manage the expansion of European enslavement and the colonization of other peoples, the definitions, criteria, and boundary lines that determine racial categories have constantly shifted over the course of U.S. history" [12]. However, the biologization of race as a system of power "has been part and parcel of racism" [10]. This is not to say, though, that race and racism are at all not real in the social sense; "[w]hile race is not imaginary—it is a very real way our society categorizes people—its intrinsic origin in biology is. Race is not an illusion. Rather, the belief in intrinsic racial difference is" [12]. The effects of these systems of oppression on physiology, too, are entirely real. Rather than

emanating from biological underpinnings, "race stands as a proxy for sociocultural, economic, and particular historical processes and experiences... while the experience of a racialized life may affect health outcomes, the concept of race itself has no biological or genetic basis" [8].

As we see is often the case with sex/gender, ethnicity is sometimes articulated to be the cultural counterpart of race, and just as frequently sloppily interchanged in attempts at political correctness. A 2001 introductory sociology textbook describes the relationship as such: "One involves traits that are biological; the other, cultural... People can fairly easily modify their ethnicity... Assuming people mate with others like themselves, however, racial distinctiveness persists over generations" [9]. As is evident in the above quote, this social-biological boundary work leads to a seemingly inevitably biologized definition of race. As Morning writes, in an analysis of various college-level textbooks, the "sociology textbooks [in her sample] suggest to students that race is a reflection, albeit unfaithful, of real underlying physical difference." Even with this distinction seemingly established, though, the use of these terms in practice employs this boundary with much less clarity; "the term ethnicity is frequently used, even when the groups in question are labelled with traditionally racial identifiers." Further, she writes that these textbooks "seem to have simply borrowed the term ethnicity to replace the word race... In other words, the concepts of race and ethnicity are interchangeable." The employment of these concepts in biology textbooks is no more thoughtful, either; "biology textbooks present definitions of race that are decidedly essentialist... the biology texts ground difference firmly at the genetic level; neither human perception nor social processes play a role" [10].

This is not to say that this sloppiness is specific to textbooks, or further that the biologization of race is at all antiquated in scientific research. In interviews with many university professors, Morning found "the most frequent definition of race among biologists was one that treated race as a biological characteristic." Still, though, the great variability in the these responses "strongly refute[s] the claim that scientists have arrived at a consensus about the nature of race" [10]. The pursuit of personalized medicine, too, has revitalized the biological

race concept as an ideologically-neutral necessity for the betterment of medical treatment outcomes. As Roberts writes, though, "[p]redicting drug response based on a patient's race rather than on genetic traits, says Lawrence LESCO of the FDA's Center for Drug Evaluation Research, is 'like telling time with a sundial instead of looking at a Rolex watch'" [12]. The endeavor to identify subpopulations with greater likelihood of responding positively to specific medical treatments is by all means a noble cause—supposing that clinically relevant differences will fall neatly along racial (read: social) divisions, though, creates a proxy under which modern racial psuedoscience can be perpetuated without critique.

Altogether, then, race/ethnicity is ill-defined both as a categorization system and, by extension, a biological mechanism. At the same time, the staying power of biologized race, especially among scientists, remains striking.

2.3 Hypotheses

As a result of the discussion above, describing the boundary work that scientists partake in to delineate between social and biological phenomena, I propose the first hypothesis:

Hypothesis 1. Of data purporting to measure sex/gender or race/ethnicity effects, data for use in biological contexts will be more likely to name such columns "Sex" or "Race" rather than "Gender" or "Ethnicity," respectively, than data collected for other purposes.

However, as argued above, despite partaking in this boundary work, positivist science often fails to integrate this social-biological delineation into measurements, data collection, and arguments for causal mechanisms. This leads to the next hypothesis:

Hypothesis 2. The distribution of entries in columns measuring sex/gender or race/ethnicity effects will be the same, regardless of which term is used to describe the column.

The operationalization of these hypotheses is described in the next section.

3 Data & Methods

This study makes use of data scraped from an online repository of open-source statistical software. All scraping and analysis scripts are freely availably online.[1]

[1]Source code is available at: *https://github.com/simonpcouch/social_divisions_in_data*

3.1 Data Acquisition

The Comprehensive R Archive Network (CRAN) is a repository of, at the time of writing, over 15,300 open-source statistical software packages for use with the R programming language. In addition to housing code libraries, many of these packages contain data, either serving as the primary purpose of the package or for use in examples showcasing the functionality of the code libraries. Those working with the R language use these libraries in a large variety of contexts, from academic research in biology, social science, and statistics, to commercial applications in finance and data science, to teaching statistics and analytics [6]. I make use of this repository in order to sample datasets used in a variety of contexts—namely, identifying those that are used for biological purposes versus those that are not. I use an automated, algorithmic approach to gather data from CRAN.

Initially, I scraped a list of all packages and relevant metadata such as package descriptions and documentation URLs from CRAN's website logs.

While CRAN offers a large sample of diverse datasets, and hosts a variety of metadata relevant to this analysis, the repository currently does not systematically classify packages based on their purpose or intent. To address this, I use a two-part approach to infer the purpose of the package. Initially, I utilize CRAN Task Views (CTV), a CRAN-hosted service offering lists of packages curated for carrying out specific tasks (e.g. "TeachingStatistics" or "ClinicalTrials") [17]. I first make use of this service to identify an initial sample of packages in each of the relevant categories. Next, I make use of package metadata to more coarsely infer the purpose of each package. Using narrative description data, packages are sorted using the presence of keywords and phrases. In combining these two data sources, I prioritize the curated CTV classifications—if a package has a classification from CTV, the package is sorted using this grouping. Then, classifiers for the remaining packages are interpolated using the keyword matching procedure. The numbers of packages in each of these groups are shown in Table 1; for computability, I take a sample of packages from the "Other" category.

For each package in the sample, the program checks whether the package contains

datasets. If it does, it stores the names of each of them. Then, for each dataset in the package, the program searches for matches to several keywords in the column names of the dataset (e.g. "Sex" or "Gender.") If any of these keywords are matched, an algorithm chooses the column most likely to contain information relevant to the keyword of interest. Finally, the program extracts all unique entries in the column (e.g. a column titled "Sex" might contain the unique entries "Female" and "Male,") and store them in a dataset giving the unique value, the number of times it appeared, the column it appeared as an entry to, the name of the dataset containing the column, and the package containing the dataset. This process is iterated over every keyword, in every column, in every dataset, in every package in the sample. Then, a set of criteria generates a cleaned version of the entries in order to allow for basic text analytics, collapsing values that encode essentially the same value (e.g. "woman", "Woman", and "W" are all encoded as "Woman.") Summary statistics about this dataset are presented in Tables 1 and 2.

3.2 Analysis

Making use of this data, I develop a set of sub-hypotheses to empirically test the hypotheses given in Section 2.3.

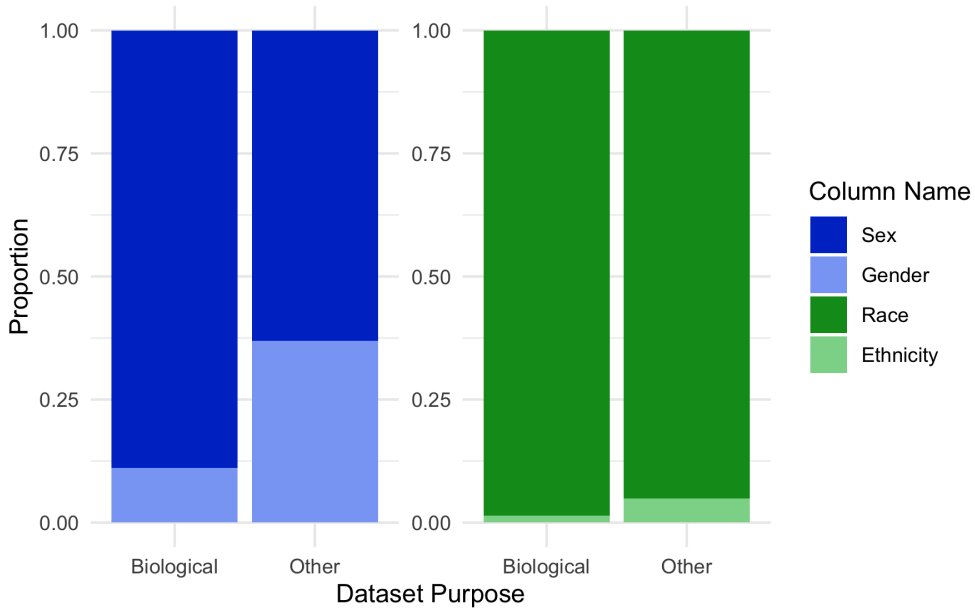
- *Hypothesis 1(a).* Datasets from packages intended for biological purposes will be more likely to refer to sex/gender measures as sex rather than gender, if one or the other is included, than datasets from packages intended for other purposes.

- *Hypothesis 1(b).* Datasets from packages intended for biological purposes will be more likely to refer to race/ethnicity measures as race rather than ethnicity, if one or the other is included, than datasets from packages intended for other purposes.

To test Hypotheses $1(a)$ and $1(b)$, I use a t-test for difference in proportions. To carry out this procedure to test Hypothesis $1(a)$, I first collect a list of all datasets in the sample, and whether that dataset supplies a column called "Sex," "Gender," or both. Then, I calculate the proportion of datasets for biological purposes that supply a column called "Sex," and the same

proportion for datasets that are not for biological purposes. Then, if the difference between the two (former minus the latter) is greater than zero, then packages intended for biological purposes are more likely to refer to sex than gender. The statistical significance of this finding is then tested to assess the validity of the above Hypothesis. A visual representation of the observed data relevant to these hypotheses are shown in Figure 1.

Figure 1: Column Name Choice by Dataset Purpose



The proportion of columns in datasets for either biological purposes, or some other purpose, named with terms evoking biological or social connotations. The difference in these proportions is tested in evaluating Hypotheses 1(a) and 1(b). See Table 2 for raw data.

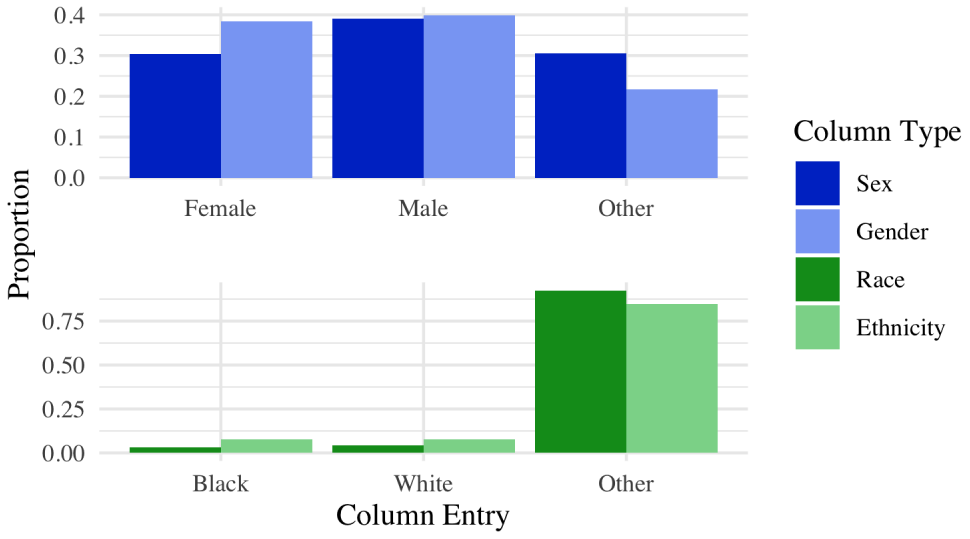
- *Hypothesis 2(a).* The distribution of values in columns named "Sex" will not differ from that of columns named "Gender." Namely, the frequencies of the entries "Male," "Female," and others (as a group) will not be different.

- *Hypothesis 2(b).* The distribution of values in columns named "Race" will not differ from that of columns named "Ethnicity." Namely, the frequencies of the entries "Black," "White," and others (as a group) will not be different.

Note that all sampled data is used to test this hypothesis, rather than just that inferred to be intended for use in biological contexts. Given the discussion above, this choice should result in greater divergence between the two distributions. This will result in a greater likelihood of statistical significance, which, in this case, leads to a more conservative evaluation of this hypothesis.

In order to test Hypotheses 2(a) and 2(b), I make use of the χ^2 (Chi-Squared) Goodness of Fit test. The Chi-Squared Goodness of Fit test can be used to test whether the observed distribution of values of a variable follows an expected distribution. In order to test this, for Hypothesis 2(a), I calculate the proportion of columns labeled "Sex" that contain at least one entry of each of "Male," "Female," or some other value. Then, this distribution is considered the expected distribution, and compared to the same distribution for values in columns labeled "Gender". The analogous procedure is used to test Hypothesis 2(b). A visual representation of the observed data relevant to these hypotheses is shown in Figure 2.

Figure 2: Proportion of Entries by Column Type



The proportion of entries in each column type allotted to the most popular entries, or some other entry. The difference in these distributions is tested in evaluating Hypotheses 2(a) and 2(b). See Tables 3 and 4 for raw data.

4 Results

Beginning with Hypothesis 1, I find that both Hypotheses are directionally supported, but to varying extents. In regard to Hypothesis 1(a), I find that datasets from packages inferred to be intended for biological research are 25.8% more likely to refer to their columns measuring sex/gender effects as "Sex" rather than "Gender" than datasets from packages inferred to be intended for other purposes. This difference is statistically significant ($p < 0.001$). As for Hypothesis 1(b), datasets from packages inferred to be intended for biological research are 3.5% more likely to refer to their columns measuring race/ethnicity effects as "Race" rather than "Ethnicity" than datasets from packages inferred to be intended for other purposes, a difference which is not statistically significant ($p = 0.110$). While I do not find statistically significant evidence for Hypothesis 1(b), I argue that not only is the finding in regard to Hypothesis 1(a) statistically significant, but is also practically significant; a difference of 25.8% represents substantial evidence of the boundary work of claims-making of "purely" biological phenomena carried out by biologists in construction of datasets.

This difference, of course, could be a practical consequence of biologists actually measuring different quantities. Thus, moving on to Hypothesis 2, I test the difference in distributions of the most common entries in sex/gender and race/ethnicity columns. Starting with Hypothesis 2(a) I find that the distribution of entries in columns named "Sex" are statistically significantly different from those named "Race" ($p = .007$, $\hat{\chi^2} = 10.0$, $df = 2$). However, I argue that this difference is not practically significant. As shown in Figure 2, "Female" and "Male" are the most common entries, by far, in columns labeled either "Sex" or "Gender." While the proportion of entries labeled "Male" is nearly identical in both columns, the main difference in distributions increasing the magnitude of the test statistic is the lesser representation of females in columns labeled "Sex." The fact, though, that terms traditionally describing gender identities like "Man," "Woman," or "Nonbinary" make up almost none of the entries in columns labelled "Gender," (2, 2, and 1 of 208 entries in the sample, respectively,) suggests that sex and gender are regarded as interchangeable terms for otherwise identical

phenomena in constructing data across the entire sample. In regard to Hypothesis 2(b), as Figure 2 reflects, the text mining procedure used in this study did not adequately capture any shared common entries in columns labeled as "Race" or "Ethnicity." As a result, the resulting counts (shown in Table 4) do not satisfy the assumptions of the χ^2 goodness-of-fit test, and it is thus inappropriate to carry out significance tests on the observed data.

5 Discussion

Support for the proposed hypotheses is partial. In the case of sex/gender, I find that datasets for use in biology are statistically and practically significantly more likely to refer to columns measuring sex/gender effects as "Sex," but further that the entries within these columns are also statistically significant. However, I argue that, in regard to Hypothesis 2(a), this finding is not practically significant. The fact that the distributions have any number of entries in common should be striking, let alone the similarity in distribution shown in Figure 2. For these reasons, I argue that this claims-making on the part of biologists, as supported by Hypothesis 1(a), is purely symbolic—sex and gender are viewed as interchangeable terms for the same phenomena, (more exactly, for sex,) and biologists are more likely to use the term "Sex" in order to suggest a more purely biological mechanism for effects observed in the data.

However, in regard to race/ethnicity, my findings are limited. While data for use in biological contexts was likely to refer to columns measuring race/ethnicity effects as "Race" than other data, this difference is not statistically significant. Further, the sample size for common entries measuring race/ethnicity effects did not meet the assumptions of the χ^2 goodness-of-fit test, so Hypothesis 2(b) remains untested. Some limitations to the methods used in this study contributing to this result are described below.

Principally, every decision-making procedure integrated into the data-scraping program is associated with some measure of error. For one, the inference on a packages purpose/intent is coarse and approximate. Based on the presence of keywords such as "genome" or "epidem" in package names and descriptions, the procedure infers that a package is likely to be used

in biological contexts. I argue, though, that misclassification error in this instance would result in smaller effect sizes, and thus decrease the likelihood of statistical significance. In this way, the inference procedure produces results that are overly conservative. Moving on, this procedure rests on the assumption that, if CRAN users are measuring sex/gender or race/ethnicity effects in their data, they would name their columns as such. Again, I argue that the violation of this assumption does not result in the exacerbation of observed effects—if these columns are named imprecisely, then the entries within them should be similarly imprecisely coded. Altogether, then, I argue that the errors associated with each of these procedures results in, in reference to Hypothesis 1, overly conservative estimates of effects. In regard to Hypothesis 2, the entry "cleaning" procedure consolidates values that are inferred to be equivalent. Most of this consolidation is trivial (e.g. "male" to "Male.") In some cases, though, these decisions require more nuance; most notably, in columns named "Sex" and "Gender" with binary entries "0" or "1," I recode these entries as "Female" and "Male," respectively, reflecting the common practice of encoding males as "successes" in binary variables. Though I manually checked that this procedure did not misclassify any entries in my initial sample (either switching "Male" and "Female" or subsituting some other value entirely,) this manual validation procedure was not automated (and thus not immediately scrutinizable or reproducible.) In the case of tests relevant to Hypothesis 2, then, excessive error results in the assumptions of the χ^2 goodness-of-fit test to be violated, as insufficiently large proportions of the sample are left "unbinned," and thus assessment of significance to be inappropriate (as was the case with testing Hypothesis 2(b).)

In addition to quantifying and scrutinizing the error associated with the procedures outlined above, future work utilizing similar methodology can take some other steps in order to examine these claims more thoroughly. Most notably, a larger sample would allow for rigorous testing of statistical significance of Hypothesis 2(a). In general, too, consideration of social divisions beyond the four examined in this paper would allow for more nuanced understanding of the ways that these categories are argued in data in general. At the same

time, these four social categories were chosen, among other reasons, for their prominence as encoded variables—for variables that are less likely to be measured and encoded in data, inversely proportionally larger samples will be necessary for valid inference. Further, the binning of entries that did not match either of the two most common entries in these columns ignores a significant portion of the structure of this data; the development of methods to examine these "unmatched" entries, and the relationships between them, in a more nuanced way could provide significant insight into the claims-making evident in these datasets. Lastly, future work should incorporate time into analyses in order to better understand how the way that these conceptualizations are encoded has changed over time.

Altogether, I have attempted to show that datasets are sites of argumentation, claims-making, and boundary work that reflect social conceptualizations held by those constructing the data.

References

[1] Alan Booth, Douglas A Granger, Allan Mazur, and Katie T Kivlighan. Testosterone and social behavior. *Social Forces*, 85(1):167–191, 2006.

[2] Patricia Hill Collins. Moving beyond gender: Intersectionality and scientific knowledge. *Revisioning gender*, pages 261–284, 1999.

[3] Anne Fausto-Sterling. The bare bones of sex: part 1—sex and gender. *Signs: Journal of Women in Culture and Society*, 30(2):1491–1527, 2005.

[4] Kanae Haneishi, Andrew C Fry, Christopher A Moore, Brian K Schilling, Yuhua Li, and Mary D Fry. Cortisol and stress responses during a game and practice in female collegiate soccer players. *Journal of strength and conditioning research*, 21(2):583, 2007.

[5] Olena Hankivsky, Kristen W. Springer, and Gemma Hunting. Beyond sex and gender difference in funding and reporting of health research. *Research Integrity and Peer Review*, 3(1):6, 2018.

[6] Kurt Hornik. The Comprehensive R Archive Network. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4):394–398, 2012.

[7] Rebecca Jordan-Young and Raffaella I Rumiati. Hardwired for sexism? approaches to sex/gender in neuroscience. *Neuroethics*, 5(3):305–315, 2012.

[8] Catherine Lee. "race" and "ethnicity" in biomedical research: how do scientists construct and explain differences in health? *Social Science & Medicine*, 68(6):1183–1190, 2009.

[9] John Macionis. Sociology. newjersey, 2001.

[10] Ann Morning. *The nature of race: How scientists think and teach about human difference*. Univ of California Press, 2011.

[11] Sarah S Richardson. *Sex itself: The search for male and female in the human genome.* University of Chicago Press, 2013.

[12] Dorothy Roberts. *Fatal invention: How science, politics, and big business re-create race in the twenty-first century.* New Press/ORIM, 2011.

[13] Rae Silver and Rebecca Jordan-Young. Sexes, genders, and brains: four scientists, four perspectives. 2014.

[14] Kristen W Springer, Jeanne Mager Stellman, and Rebecca M Jordan-Young. Beyond a catalogue of differences: a theoretical frame and good practice guidelines for researching sex/gender in human health. *Social science & medicine*, 74(11):1817–1824, 2012.

[15] Ethel Ed Tobach and Betty Ed Rosoff. *Challenging racism and sexism: Alternatives to genetic explanations.* Feminist Press at the City University of New York, 1994.

[16] Sari M Van Anders and Neil V Watson. Social neuroendocrinology. *Human nature*, 17(2):212–237, 2006.

[17] Achim Zeileis. CRAN Task Views. *R News*, 5(1):39–40, 2005.

A Appendix

Table 1: Number of Packages by Group

Purpose	Census Size	Sample Size
Biological Research	925	925
Other	14383	1694

Counts of packages by intended purpose. See Section 3.1 for discussion of the data acquisition and sampling process.

Table 2: Number of Flagged Columns by Group

Purpose	Sex	Gender	Race	Ethnicity
Biological Research	72	9	135	2
Other	419	205	468	67

Counts of relevant columns in the sample by column name and intended purpose of the package.

Table 3: Number of Unique Entries by Purpose & Column Type

Entry	Sex		Gender	
Female	25	100	2	78
Male	34	127	2	81
Other	13	112	5	40

Counts of common unique entries by column name (either sex or gender) and dataset purpose, where values in gray columns represent counts from datasets intended for biological research, while values in white columns represent counts from datasets intended for other purposes.

Table 4: Number of Unique Entries by Purpose & Column Type

Entry	Race		Ethnicity	
Black	0	20	0	2
Other	133	420	2	20
White	2	23	0	2

Counts of common unique entries by column name (either race or ethnicity) and dataset purpose, where values in gray columns represent counts from datasets intended for biological research, while values in white columns represent counts from datasets intended for other purposes. See Section 4 for discussion of the significance of these lower counts.