FAIR2: A framework for addressing discrimination bias in social data science

Francisca Garcia-Cobián Richter[1], **Emily Nelson**[1], **Nicole Coury**[1], **Laura Bruckman**[1], **Shanina Knighton**[1,2]

[1]Case Western Reserve University; [2]Center for Infection Prevention and Control Research

Abstract

Building upon the FAIR principles of (meta)data (Findable, Accessible, Interoperable and Reusable) and drawing from research in the social, health, and data sciences, we propose a framework -FAIR2 (Frame, Articulate, Identify, Report) - for identifying and addressing discrimination bias in social data science. We illustrate how FAIR2 enriches data science with experiential knowledge, clarifies assumptions about discrimination with causal graphs and systematically analyzes sources of bias in the data, leading to a more ethical use of data and analytics for the public interest. FAIR2 can be applied in the classroom to prepare a new and diverse generation of data scientists. In this era of big data and advanced analytics, we argue that without an explicit framework to identify and address discrimination bias, data science will not realize its potential of advancing social justice.

Keywords: *Discrimination Bias; Social Data Science Framework; Experiential Knowledge; Causal Diagrams.*

1. Introduction

There is a long, damaging history of purportedly objective use of data that has contributed to reinforcing stereotypes driving fear, isolation, and discrimination in society. Social and health science scholarship of the early 1900s in the United States had an unfortunate role in establishing false connections between crime and the social construct of race (Muhammad, 2019). Prominent social scientists and statisticians used Census Bureau and prison data to make flawed causal claims linking race to crime, helping to cement discrimination in all aspects of society. Mathematician Kelly Miller, sociologist W. E. B. DuBois, and journalist Ida B. Wells were among the Black scholars and activists who counteracted these narratives with carefully crafted arguments hinged on data, logic, and domain-expert knowledge. However, the academic community mostly dismissed this work. Only in 2020 have academic associations issued statements recognizing their lack of understanding of racism and its impact on their work -if not apologizing for the harms caused (American Economic Association, 2020).

In the current era of big data and data technologies, discrimination, reflected in social data in a multiplicity of ways, is a main source of bias. Bias here is generally defined as distortions in inference stemming from data or assumptions. Yet until recently, discrimination in data -particularly discrimination based on the social construct of race- has mostly been ignored by the academic community. Consequently, analyses using these data have little ability to address discrimination or worse yet, may strengthen biased beliefs and perpetuate discrimination. While data analysts may agree that discrimination in society manifests itself in social data, we argue that without an explicit framework to identify and address this problem, data science will not realize its potential of supporting social justice.

In this article, we offer one such framework -FAIR2- that draws from recent literature in the social and health sciences, data science, computer science, community-engaged research practices and metadata principles. It complements the ethical standards of FAIRification (Findable, Accessible, Interoperable and Reusable) principles (Wilkinson et al., 2016) with a set of four additional principles (Frame, Articulate, Identify, Report) specific to working with social data for social impact. Primarily developed to guide students doing data science for social impact, FAIR2 can be more broadly applied to foster stronger communication between researchers, practitioners, and community members whose experiences are represented in the data. Section 2 introduces the FAIR2 framework. Section 3 illustrates the use of FAIR2 with an example of data analytics in the area of public assistance programs. Section 4 provides concluding thoughts.

2. FAIR2

Since the FAIR data principles were proposed in 2016 by a diverse group of stakeholders in academia and industry, their adoption continues to grow. FAIR stands for Findability, Accessibility, Interoperability, and Reusability principles, meant to ease the ability of machines and humans to make informed use of data and maximize the value added of data analytics in a transparent, ethical way (Rocca-Serra et al., 2022). Building on FAIR, we propose an additional set of principles pertinent to social data and analytics, FAIR2. The FAIR2 initials refer to: Frame, Articulate, Identify and Report. This framework recognizes that *data do not speak for themselves*, that observational data reflect discrimination in society, and that an explicit framework to identify and address discrimination biases will further data science's potential to advance social justice.

2.1. Frame

Frame metadata and data with historical context and experiential knowledge of those represented in the data. The inclusion of individuals in administrative data in healthcare, homelessness, policing, among others, is influenced by discrimination. Drawing from the experiential knowledge of people intersecting with these systems will enrich the understanding of who is represented and not represented in the data, how reliable the data are, and what can be learned from it. The Human Rights-Based Approach to Data (OHCHR, 2018) posits that participation by relevant populations in data collection and analysis is key to enhancing the use of data in alignment with international human rights norms. When data has been collected by administrative systems or machines, community participation in establishing the metadata takes on a heightened role. Integrating community knowledge into the metadata can be accomplished via the inclusion of qualitative literature and through designed collaboration meetings -Data Chats- with community members. Data Chats are created with the intention of "center[ing] residents' knowledge, community understanding, and experiences as much as quantitative data (Cohen, Rohan, Pritchard, & Pettit, 2022)". Their structure allows researchers to collaborate and learn from community members while sharing information derived from data. Implementation considerations include developing informed consent forms, planning for accessibility of meeting space and time, sharing a meal, all of which reflects respect and appreciation towards community collaborators. This approach ties in well with the goals of the FAIR2 framework.

2.2. Articulate

Articulate the general model as a causal graph to explicitly state model assumptions (background knowledge) and hypotheses about the role of discrimination in the social phenomenon studied. It has been said that the logic of inference in policy analysis can be summarized as "assumptions + data → conclusions" (Manski, 2013). In other words, the data do not speak for themselves; assumptions can be a source of subjectivity in inference.

Directed Acyclic Graphs (DAGs) are a collection of nodes and directed edges connected under certain conditions to represent a causal model (Pearl, 1995). DAGs make explicit the assumptions embedded in the model, helping to clarify the source of these assumptions (what knowledge and whose knowledge?) and their implications for model estimates (Pearl & Mackenzie, 2018). They can represent the larger context underlying the social phenomenon of interest and the sources of discrimination in outcomes. Furthermore, DAGs facilitate the awareness of selection bias and collider bias that can interfere with the identification of causal effects and interpretation of predictive algorithms. There is much to learn from the recent work of researchers across multiple social science disciplines that have used DAGs to clarify the role of discrimination -in particular racism- on outcomes of well-being (Howe, Bailey, Raifman, & Jackson, 2022), to improve the performance of machine learning (Robinson, Renson, & Naimi, 2020), and to advances in algorithmic fairness (Kilbertus et al., 2017).

2.3. Identify

Identify bias embedded in the data and variables of interest, aiming to minimize bias and report on limitations due to bias. Here we draw from the work of (Kleinberg, Ludwig, Mullainathan, & Sunstein, 2018) and (Lundberg, Johnson, & Stewart, 2021) to systematically analyze potential biases in the estimand and choice of variables used in the model. We set out to answer the following questions: (1) What is the unit-specific quantity of interest -USQ- and the target population? (2) What biases may be introduced by each variable and by using measured versus desired variables? (3) What is the role of variables representing the sensitive attributes (subject to discrimination) in the model? (4) For whom is the USQ missing in the data (selective labels problem; collider bias), why, and how will this affect model estimates?

2.4. Report

Share findings and seek feedback from members or agencies in the community who have experiential knowledge of the social issue analyzed. The increased use of mixed methods research and intersectionality theory (Abrams, Tabaac, Jung, & Else-Quest, 2020) reveal the growing popularity of context-rich and collaborative data. Research has shown that openly communicating and engaging with communities to disseminate data can build trust and ground academic datasets in real world issues (Schalet, Tropp, & Troy, 2020). Sharing data improves health equity, as has been shown in harm reduction strategies for people who use drugs (Salazar, Vincent, Figgatt, Gilbert, & Dasgupta, 2021), and with emergency housing programs (Lane, McClendon, & Matthews, 2017). Engaging community stakeholders in analyzing and reporting findings completes the circuit of utilizing data without losing its human context.

3. Application to an Analysis of Nutrition Assistance Recertification

We illustrate the use of the FAIR2 framework with an analysis of recertification in the Supplemental Nutrition Assistance Program (SNAP), one of the main public assistance programs in the United States. SNAP's predecessor, the Food Stamps Program, was developed in 1933 during the Great Depression to support farmers and people facing food insecurity. Today, SNAP is offered as a food voucher via Electronic Transfer Cards, conditional on meeting low-income thresholds that vary by states. SNAP is among the strongest programs in the US social safety net, with a countercyclical multiplier effect during economic downturns (Canning & Stacy, 2019). States require individuals to follow a recertification process every 6 to 12 months involving detailed proof of income and an interview. Research suggests that up to half of beneficiaries who exit SNAP within their first year were still eligible (Gray, 2019). Failing to recertify despite qualifying -here denoted as FR- is costly to individuals and administrative agencies. *Researchers have sought to study who is affected by FR and what can be done to reduce the rate of FR.*

Frame - Data Chat with SNAP participants who have experienced the recertification process call attention to (1) an under-resourced system: hours long wait times on the phone to make interview appointments, letters requiring additional information that come close to the interview day; (2) a stressful process that can sometimes feel confrontational and the need to "suck up your pride"; (3) a system that seems to penalize any change in employment; (4) within- and across-locality variation in service quality, with higher income localities seemingly performing better than those with higher need and some case workers showing extreme dedication despite the difficulties of navigating an under-resourced system.

Articulate - A common approach to characterize the population experiencing FR is to use administrative data to estimate a regression model for the FR outcome, with demographic and economic characteristics (D&E) as predictors. The left DAG of Figure 1 presents a naive model including "race" and D&E variables compatible with such regression. It implies strong assumptions about the meaning of "race," represented as an individual trait and not directly related to other societal factors that impact FR. The right DAG, explained in the caption of Figure 1, embeds historical and experiential knowledge from our Data Chats and the literature, highlighting systemic issues that are relevant to inform policy. This DAG is inspired by recent work on causal diagrams for studying racial health disparities (Howe et al., 2022; Robinson et al., 2020).

Identify - (1) Let SNAP enrollees subject to recertification in locality L constitute the target population and FR -failure to recertify when eligible- the unit-specific variable of interest.
(2) Unable to learn whether enrollees meet eligibility requirements for recertification from administrative data, researchers have sought to flag FR if an individual drops from SNAP and re-enters within 1-3 months (Kenney et al., 2022). This is denoted as churn.
Acknowledging discrepancies between the theoretical and empirical estimand would point researchers to explore re-entry beyond 90 days in the administrative data or through

qualitative knowledge. It may also suggest linking administrative data to capture eligibility requirements and thus FR rather than churn episodes. (3) The misconception of "race" as a demographic variable (R in Fig. 1, left DAG) ignores important plausible pathways by which discrimination may be directly impacting FR (HP → CP→ S or FR in Fig. 1, right DAG). While it is valuable to assess differences in FR by racialized groups, acknowledging the underlying mechanisms that lead to inequities in outcomes have strong implications for setting up research designs and identifying policy solutions. (4) Administrative data generated with low resources and characterizing a population facing distress are likely to exhibit non-random missingness issues. Are these patterns consistent with variations in resources or recertification requirements across demographic or geographic groups highlighted in our community conversations?

Report - Co-creating findings allows researchers to engage community members as experts on the social issues they have experienced and share power with the community. Data Chat participants review summaries and conclusions from SNAP recertification analyses and our synthesis of their initial thoughts. They incorporate their input for a final report that will be shared with local administrators of the SNAP program.

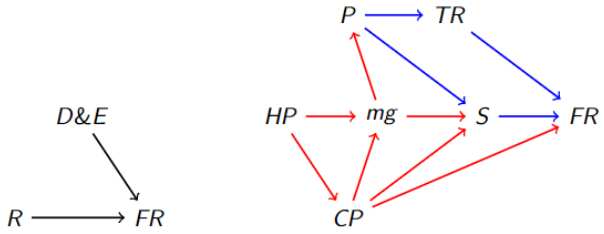

Figure 1. Directed Acyclic Graphs modeling failure to recertify for food assistance (FR). The left graph represents a naive model that aims to explain differences in FR with demographic and economic characteristics (D&E) and race (R). The right model includes historical and experiential knowledge. HP are historical processes that shaped discrimination in US society (like Jim Crow or Redlining); CP are contemporary policies that allocate scarce resources to the social safety net (weak transportation system, understaffed public assistance administration); mg is classification into marginalized grouping; P is poverty; TR is time and resource scarcity; S is stigma.

4. Concluding thoughts

Discrimination in society may influence who is represented and not represented in the data, and how. Variables of interest may be measured with less accuracy in the presence of discrimination, or the metrics used to proxy desired variables may be flawed by discrimination. Furthermore, modeling assumptions, not always explicitly stated in data analyses, can inadvertently carry biases reflective of discrimination in society. The increased

use of scoring algorithms to inform decision making in human services has heightened the need to integrate experiential community knowledge in the development of data technologies for the public interest (Roewer-Despres & Berscheid, 2020). We draw from the work herein cited to build FAIR2 as a tool to address discrimination bias in social data and analytics, strengthen education in data science for social impact, and ultimately propel the field of public interest technology to advance social justice with equity.

Acknowledgements

We thank our community collaborator Ms. Alice Jackson, all Data Chat participants, and the PIT-UN Community-Academic Advisory group at Case Western Reserve University for their insightful contributions. This work was generously funded by grant #015865 from the Public Interest Technology University Network - New America Foundation.

References

Abrams, J. A., Tabaac, A., Jung, S., & Else-Quest, N. M. (2020). Considerations for employing intersectionality in qualitative health research. Social Science & Medicine, 258, 113138. https://doi.org/10.1016/j.socscimed.2020.113138

American Economic Association. (2020, June 5). Statement from the AEA Executive Committee. Retrieved from https://www.aeaweb.org/news/member-announcements-june-5-2020

Canning, P., & Stacy, B. (2019). The Supplemental Nutrition Assistance Program (SNAP) and the Economy: New Estimates of the SNAP Multiplier (Economic Research Report No. 291963). United States Department of Agriculture, Economic Research Service. Retrieved from https://econpapers.repec.org/paper/agsuersrr/291963.htm

Cohen, M., Rohan, A., Pritchard, K., & Pettit, K. L. S. (2022). Guide to Data Chats: Convening Community Conversations about Data. Urban Institute.

Gray, C. (2019). Leaving benefits on the table: Evidence from SNAP. Journal of Public Economics, 179, 104054. https://doi.org/10.1016/j.jpubeco.2019.104054

Howe, C. J., Bailey, Z. D., Raifman, J. R., & Jackson, J. W. (2022). Recommendations for Using Causal Diagrams to Study Racial Health Disparities. American Journal of Epidemiology, 191(12), 1981–1989. https://doi.org/10.1093/aje/kwac140

Kenney, E. L., Soto, M. J., Fubini, M., Carleton, A., Lee, M., & Bleich, S. N. (2022). Simplification of Supplemental Nutrition Assistance Program Recertification Processes and Association with Uninterrupted Access to Benefits Among Participants with Young Children. JAMA, 5(9) https://doi.org/10.1001/jamanetworkopen.2022.30150

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. Advances in Neural Information Processing Systems, 30. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. Journal of Legal Analysis, 10, 113–174. https://doi.org/10.1093/jla/laz001

Manski, C. F. (2013). Public Policy in an Uncertain World: Analysis and Decisions. Cambridge, MA: Harvard University Press.

Lane, S. R., McClendon, J., & Matthews, N. (2017). Finding, Serving, and Housing the Homeless: Using Collaborative Research to Prepare Social Work Students for Research and Practice. Journal of Teaching in Social Work, 37(3), 292–306. https://doi.org/10.1080/08841233.2017.1317689

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. American Sociological Review, 86(3), 532–565. https://doi.org/10.1177/00031224211004187

Muhammad, K. G. (2019). The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America, With a New Preface. Cambridge, MA: Harvard University Press.

OHCHR. (2018). A Human Rights-Based Approach to Data. Geneva: Office of the United Nations High Commissioner for Human Rights. Retrieved from Office of the United Nations High Commissioner for Human Rights website: https://www.ohchr.org/sites/default/files/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect (1st ed.). USA: Basic Books, Inc.

Robinson, W. R., Renson, A., & Naimi, A. I. (2020). Teaching yourself about structural racism will improve your machine learning. Biostatistics, 21(2), 339–344. https://doi.org/10.1093/biostatistics/kxz040

Rocca-Serra, P., Sansone, S.-A., Gu, W., Welter, D., Abbassi Daloii, T., & Portell-Silva, L. (2022). Reflections on the Ethical values of FAIR. In D2.1 FAIR Cookbook. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6783564

Roewer-Despres, F., & Berscheid, J. (2020, November 29). Continuous Subject-in-the-Loop Integration: Centering AI on Marginalized Communities. arXiv. Retrieved from http://arxiv.org/abs/2012.01128

Salazar, Z. R., Vincent, L., Figgatt, M. C., Gilbert, M. K., & Dasgupta, N. (2021). Research led by people who use drugs: Centering the expertise of lived experience. Substance Abuse Treatment, Prevention, and Policy, 16(1), 70. https://doi.org/10.1186/s13011-021-00406-6

Schalet, A. T., Tropp, L. R., & Troy, L. M. (2020). Making Research Usable Beyond Academic Circles: A Relational Model of Public Engagement. Analyses of Social Issues and Public Policy, 20(1), 336–356. https://doi.org/10.1111/asap.12204

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18