OVERRIDING RACE AND CLASS BIAS: EQUITY IN TECHNOLOGY-MEDIATED DIAGNOSIS

By Vera K. Wilde, Ph.D.

Abstract: Racial and intersectional health disparities cost billions of dollars and tens of thousands of lives annually. Medical diagnosis decision tools—decision-making technologies that seem neutral and scientific, but require human input and interpretation—might institutionalize bias in seemingly neutral ways. This research generates evidence supporting both potentials.

Racial and intersectional (e.g., race plus class) health disparities are amply documented (National Research Council 2003; NRC 2010). This is a fiscal as well as a social justice crisis: the Institute of Medicine cites a $76 billion annual cost of racial disparities in U.S. healthcare (LaVeist, Gaskin, and Richard 2009; National Research Council 2003). Less agreement exists around what causes these disparities. Experimental and observational evidence suggests that contributing factors include differential clinician attributions, implicit bias, structural incentives, patient subcultural preferences, perceived discrimination, and stereotype threat (A. R. Green et al. 2007; Brownlee 2008; Ayanian et al. 1999; Byrne et al. 2011; Williams and Mohammed 2008; Trierweiler et al. 2000; Strakowski et al. 1996).

Psychiatric misdiagnosis of African-Americans with low socio-economic status might be particularly common, because blacks' mental health has been politicized since the antebellum era (Harris-Perry 2011). Rather than receding, this politicization arguably escalated in the mid-late 20th century. During the Civil Rights era, police and doctors came to view politically and socially active black men as paranoid and dangerous (Metzl 2009). The psychiatric discourse reframed schizophrenia, previously a disease of disorganized and disobedient housewives, in terms of racialized aggression. Today, schizophrenia is disproportionately diagnosed in poor racial minorities, especially migrant communities from developing countries, across the U.S., Britain, Western Europe, and Israel (Bresnahan et al. 2007; Selten and Cantor-Graae 2005; Littlewood and Lipsedge 1992). This disproportionate diagnosis might reflect disparities in a range of biopsychosocial variables that could contribute to disparate disease base rates among racial subgroups.

More broadly, medical diagnosis is significant because missed diagnoses involving "a principal underlying disease or primary cause of death" occur at an estimated rate of 8-24% in contemporary U.S. hospitals (Shojania 2003). Adverse events and serious errors are also common and often result from misdiagnosis (Rothschild et al. 2005). Misdiagnosis is common, costly, and disproportionately affects historically disadvantaged groups.

Luckily, medical diagnosis decision tools are increasingly used to assist in the diagnosis process (Bond, Schwartz, Weaver, Levick, Giuliano, and Graber 2010; Graber, Tompkins, and Holland 2009). These tools have the potential to advance diagnosis speed and accuracy—or to hinder that same accuracy within some subsets by institutionalizing bias in neutral-seeming ways. My research finds evidence supporting both potentials.

In a randomized controlled survey experiment, the potential for technology-mediated bias—bias in decisions that appear neutral and scientific, but could actually institutionalize prejudice—comes out in an order effect. In the experiment, people read a patient vignette and generated a probability-ranked differential diagnosis by entering patient symptoms into a leading Bayesian-updating diagnosis decision support tool. Results indicate people who see an African-American patient with welfare/disability status are significantly more likely to generate a first-ranked diagnosis that is psychiatric in nature. This order effect probably results from racial stereotypes rendering psychiatric symptoms more cognitively accessible from the same medical case vignette for this subset of patients, resulting in ordering changes in symptom input. It is consistent with what we know about disproportionate diagnosis of psychiatric illness in African-Americans in the U.S. and analogous minority groups in other countries. Future research might compare ways of potentially correcting for this type of effect.

The same survey experiment also suggests that the disparate order effect did not extend downstream to the user's evaluation of the correct diagnosis. So while people who saw an African-American patient on welfare/disability were more likely to generate a first diagnosis that was psychiatric using the diagnosis tool, they were no less likely to then believe the correct diagnosis was psychiatric or to make the correct diagnosis. This was in a difficult case—Huntingdon's—where psychiatric misdiagnosis is plausible.

The experiment also tested for potential stereotype threat, or the possibility that priming group identity might activate negative stereotypes about women and African-Americans, impairing these groups' effectiveness in using technology-mediated diagnosis tools (Danaher, Kelly, and Crandall 2008; Nguyen and Ryan 2008; Steele and Aronson 1995; Steele, Spencer, and Aronson 2002). Stereotype threat results were null, supporting the potential of diagnosis decision support tools such as this one to help equalize evidence-based healthcare access across groups when the end users are medical non-professionals such as patients themselves. The rest of this summary presents the experimental design and results in greater detail.

Against the Grain

Overall, this research both builds on and challenges existing social and cognitive psychology, political science, and sociology research on intergroup prejudice, attribution bias, and intersectional (here, race plus class) bias. Leading theories of implicit and explicit racial attitudes predict that racial bias against African-Americans and particularly against negative stereotype-conforming African-Americans is pervasive, harmful to minorities, and affects decisions—especially under conditions that should contribute to automatic as opposed to controlled cognitive processing, such as ambiguity and the ability to attribute decisions to factors other than prejudice (Allport 1979; Bargh 1994; Bertrand Mullainathan 2004; Devine, Plant, Amodio, Harmon-Jones, and Vance 2002; Dovidio, Glick, and Rudman 2005; Peffley, Hurwitz, and Sniderman 1997; Pettigrew 1979; and Sidanius and Pratto 2001). By contrast, I compile novel evidence suggesting that racial and intersectional bias do not systematically characterize technology-mediated decisions.

As numerous observational and experimental studies would predict, my research shows that African-Americans on welfare/disability appear to be more vulnerable to potential psychiatric misdiagnosis, as reflected in a statistically significantly higher probability of the first diagnosis the decision tool lists being psychiatric. However, this order effect does not appear to have practical consequences for diagnostic accuracy. So the good news is that regular people can get the right diagnosis using a diagnosis decision support tool—even in a particularly difficult-to-diagnose case, Huntingdon's, in a group we would expect to be particularly vulnerable to bias in diagnosis, African-Americans who appear to be on welfare/disability.

Survey Experimental Design

The survey experiment was conducted online between Feb. 13, 2012, and Oct. 6, 2013. The online administration mode avoided contamination from possible, unintended race and gender effects from an individual survey administrator (participant-observer effects; e.g., Gosling et al. 2004; Orne 2009; Osborne 2001). Participants were workers on Amazon.com's Mechanical Turk (MTurk) platform located in the U.S. MTurk is an Internet survey platform that facilitates simple, inexpensive recruitment and payment of subjects. Individuals undertake "Human Intelligence Tasks" (HITs) on MTurk, which can include survey experiments. MTurk data compare favorably with typical experimental political science and psychology data—which are collected using local convenience or student samples—in terms of internal and external validity. Research indicates MTurk samples replicate diverse experimental findings, even though their characteristics differ from the general population's (Berinsky, Huber, and Lenz 2012; Buhrmester, Kwang, and Gosling 2011; Gosling et al. 2004) . Moreover, replication and triangulation, not random selection, establish generalizability of experimental findings. These results replicate across survey experiments using other technology-mediated decision support tools, as well as across diverse data sources, in other research (Wilde 2014). Thus, it is reasonable to assume that the inferential statistics obtained from this series of experiments tend to be low in bias.

The survey experimental design randomly assigned participants to different race and background information variable value conditions. Participants were randomly assigned to view a vignette consisting of relevant text alongside a photo that conveyed the race of the mock subject (here, the ostensible patient—and in parallel survey experiments that replicated results in other technology-mediated decision tool contexts, the polygraph subject or food stamp applicant). The photos were normed along relevant dimensions: age, familiarity, mood, memorability, and picture quality (Kennedy, Hope, and Raz 2009). Photos come from the Center for Vital Longevity Face Database. They feature neutral facial expressions and gray backgrounds (Minear and D. C. Park 2004).

All materials were pretested in Charlottesville, Virginia between January 2012 and September 2013, and are available in the University Library (LIBRA) repository and online. All recorded observations were utilized in the reported analyses, with the following pre-determined exclusions: (1) non-U.S. respondent location according to IP address, (2) repeated study completions from the same IP address, (3) non-compliance with quality control measures, established through automated checks, and (4) failure to enter a valid response ID (a code all respondents received at the end of the survey) when prompted.

The primary dependent variable of interest is the diagnosis generated for a patient through use of a diagnosis decision tool. Its values are whether participants generated the correct diagnosis using the tool (valued at 0 for no, 1 for yes), whether the first diagnosis the tool listed given the symptoms participants entered for the patient was psychiatric in nature (valued at 0 for no, 1 for yes), and whether participants thought the correct diagnosis was psychiatric in nature (valued at 0 for no, 1 for yes). Operationalizing diagnostic accuracy through this array of objective and subjective, process and outcome-oriented measures strengthens the outcome variable's construct validity.

Assessing whether the first diagnosis the tool listed was psychiatric in nature was a particularly tough test of the intersectional bias hypothesis, that an African-American patient on welfare/disability presenting with symptoms of a difficult diagnosis (Huntingdon's) would be assessed with the same degree of accuracy as a white patient not on welfare/disability presenting with the same symptoms.

The survey instrument in this experiment has three parts: the medical puzzle, consisting of the patient scenario and online diagnosis tool; post-test survey questions on patient assessments and user experience with the tool; and pre- or post-test demographic questions. In the first part, the race variable values were black or white, operationalized through photographs embedded in an experimental vignette. The faces were morphed to be slightly more similar. The background information variable values are being on welfare and having applied for disability, or not. The text operationalizing this variable in the welfare/disability condition read:

> He had provided an emergency contact, and gave the doctor permission to call her. The emergency contact was his social worker. She didn't know anything about his medical history, except that he recently applied for disability. She had referred him to a service organization for assistance completing the application due to his poor organization.

In the non-welfare/disability condition, the same portion of the text instead read:

> … The emergency contact was his neighbor. She didn't know anything about his medical history. She sometimes helped him with paperwork due to his poor organization.

The welfare/disability treatments systematically covaried to maximize the strength of tests for disconfirming evidence. This is realistic and internally consistent, because a poor person who is ill would have a greater need to apply for disability than a person who does not financially need to work. This covariation also generates a strong test of the hypothesis that racial and class biases compound, since there are particular substereotypes about lazy, undeserving blacks seeking public assistance (Gilens 2000; Mendelberg 2001, 2008). Being on welfare and applying for disability can both be conceived as public assistance-seeking behaviors.

This vignette appears in the experiment as a case description in the medical puzzle which participants solve using the diagnosis tool. The vignette is an excerpt of a published case report written for a lay audience, minimally edited for length, clarity, and the experimental manipulation detailed above (Sanders 2005). This origin further bolsters results' external validity, because the case represented was a real one, and the description was written by a relevant expert.

Overall, these independent variable operationalizations have good mundane realism—they are reasonably true-to-life in the sense that healthcare workers see patients' race and often know what insurance they have, clueing them into patients' welfare/disability status. When treating a very sick patient, they frequently call the emergency contact. Patients routinely answer the same demographic questions asked in this study.

Results

Patient race and welfare/disability status and their interaction do not significantly add to the explanation of variability in dependent variable values relating to whether or not the correct diagnosis was produced using the medical diagnosis tool. Since it involved a wide variety of diagnoses, the data from which these binary dependent variables were constructed was more complex than in the other case studies of administrative technologies. Consequently, it makes sense to triangulate a few different constructions of a binary outcome from these data.

Medical Diagnosis – Characteristics of Diagnoses, (SE)			
	Correct Diagnosis	First Diagnosis Psychiatric	Correct Diagnosis Psychiatric
African-American	-0.132 (0.333)	-0.360 (0.183)	0.066 (0.167)
Welfare/disability status	-0.067 (0.324)	-0.254 (0.179)	0.197 (0.166)
African-American X welfare/disability status	-0.360 (0.489)	**0.582*** (0.253)	-0.269 (0.232)
Constant	-2.558*** (0.227)	-0.688 (0.124)	0.055 (0.117)
N = 1204			

* = $p < 0.05$, *** = $p < 0.000$. Results reflect coefficients from logistic regression models, with standard errors in parentheses.

Results suggest increased cognitive availability of racial stereotypes, because an African-American patient on welfare/disability is significantly more likely to be given a first possible diagnosis in a list of possible diagnoses generated from the tool that is psychiatric ($p = 0.022$). The correct diagnosis in this case (Huntingdon's chorea) is physical, and psychiatric misdiagnosis in cases of physical illness can do harm through withholding needed treatment. However, the order effect does not appear to go on to affect practically meaningful outcomes including whether the correct diagnosis is obtained from the tool, and whether the correct diagnosis is deemed to be psychiatric rather than physical.

Conclusion

Large bureaucracies such as governmental and military institutions, corporations, and educational systems have long sought to routinize and render invisible discretionary power under the auspices of science—usually with unintended consequences (March and Olsen 1989; Porter 1996; Ross 1992; Scott 1998). New technologies create new opportunities for this type of centralized standardization enterprise, while also making it possible to better decentralize informed decision-making, by making expert knowledge and tools more readily available to end users—citizens, patients, and clients alike. In this way, technology has always had the dual potential of institutionalizing bias while appearing neutral on one hand—and empowering traditionally disempowered groups on the other. Medical diagnosis decision support tools are no exception to this historical rule.

Future research on medical technologies like diagnosis decision support tools might incorporate groups of subjects with varying levels of expertise, in situations with different levels of time pressure and varying correct diagnoses. It would be ideal to triangulate field observational and experimental data sources. And identifying a mechanism whereby subjects might correct for technology-mediated bias, as appears to potentially be the case for the order effect bias in the survey experiment reported here, should also be a priority.

Bibliography

Allport, Gordon W. 1979. *The Nature of Prejudice*. Unabridged, 25th anniversary ed. Reading, Mass: Addison-Wesley Pub. Co.

Author. 2014. "Neutral Competence? Polygraphy and Technology-Mediated Administrative Decisions." Doctoral dissertation.

Bargh, John A. 1994. "The Four Horsemen of Automaticity: Awareness, Efficiency, Intention, and Control in Social Cognition." In *Handbook of Social Cognition*, Vol. 2, 1:1–31. Hillsdale, NJ: Erlbaum.

Berinsky, A. J., G. A. Huber, and G. S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3) (March 2): 351–368.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4) (September): 991–1013.

Bond, William F., Linda M. Schwartz, Kevin R. Weaver, Donald Levick, Michael Giuliano, and Mark L. Graber. 2010. "A Qualitative Review of Differential Diagnosis Generators". Conference poster presentation presented at the Diagnostic Errors in Medicine Conference, October, Toronto, Ontario, Canada. http://www.isabelhealthcare.com/pdf/Bond_A_Qualitative_Review_of_Differential_Diagnosis_Generators_DEM_2010.pdf.

Bresnahan, Michaeline, Melissa D Begg, Alan Brown, Catherine Schaefer, Nancy Sohler, Beverly Insel, Leah Vella, and Ezra Susser. 2007. "Race and Risk of Schizophrenia in a US Birth Cohort: Another Example of Health Disparity?" *International Journal of Epidemiology* 36 (4) (August): 751–758.

Brownlee, Shannon. 2008. *Overtreated: Why Too Much Medicine Is Making Us Sicker and Poorer*. Pbk. ed. New York, NY: Bloomsbury.

Buhrmester, M., T. Kwang, and S. D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6 (1) (February 3): 3–5.

Byrne, Margaret M., L. Jill Halman, Leonidas G. Koniaris, Peter A. Cassileth, Joseph D. Rosenblatt, and Michael C. Cheung. 2011. "Effects of Poverty and Race on Outcomes in Acute Myeloid Leukemia:" *American Journal of Clinical Oncology* 34 (3) (June): 297–304.

Danaher, Kelly, and Christian S. Crandall. 2008. "Stereotype Threat in Applied Settings Re-Examined." *Journal of Applied Social Psychology* 38 (6): 1639–55.

Devine, Patricia G., E. Ashby Plant, David M. Amodio, Eddie Harmon-Jones, and Stephanie L. Vance. 2002. "The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond without Prejudice." *Journal of Personality and Social Psychology* 82 (5): 835–48.

Dovidio, John F., Peter Glick, and Laurie Rudman, eds. On the Nature of Prejudice: Fifty Years after Allport. 2005. Malden, MA: Blackwell Pub.

Gilens, Martin. 2000. *Why Americans Hate Welfare: Race, Media, and the Politics of Antipoverty Policy*. Studies in Communication, Media, and Public Opinion. Chicago: University Of Chicago Press.

Gosling, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver P. John. 2004. “Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires.” *American Psychologist* 59 (2): 93–104.

Graber, Mark, David Tompkins, and Joanne J. Holland. 2009. “Resources Medical Students Use to Derive a Differential Diagnosis.” *Medical Teacher: An International Journal of Education in the Health Sciences* 31 (6): 522–27.

Green, Alexander R., Dana R. Carney, Daniel J. Pallin, Long H. Ngo, Kristal L. Raymond, Lisa I. Iezzoni, and Mahzarin R. Banaji. 2007. “Implicit Bias among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients.” *Journal of General Internal Medicine* 22 (9): 1231–38.

Harris-Perry, Melissa V. 2011. *Sister Citizen: Shame, Stereotypes, and Black Women in America*. New Haven: Yale University Press.

Kennedy, K. M., Hope, K., & Raz, N. 2009. Lifespan adult faces: norms for age, familiarity, memorability, mood, and picture quality. *Experimental Aging Research*, 35(2), 268–275.

LaVeist, Thomas A., Darrel J. Gaskin, and Patrick Richard. 2009. “The Economic Burden of Health Inequalities in the United States”. Joint Center for Political and Economic Studies.

Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation.

Littlewood, Roland, and Maurice Lipsedge. 1992. “Schizophrenia among Afro-Caribbeans.” *The British Journal of Psychiatry* 160: 710–11.

March, James G., and Johan P. Olsen. 1989. *Rediscovering Institutions: The Organizational Basis of Politics*. New York: Free Press.

Mendelberg, Tali. 2001. *The Race Card : Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton N.J.: Princeton University Press.

———. 2008. “Racial Priming Revived.” *Perspectives on Politics* 6 (1): 109–23.

Metzl, Jonathan. 2009. *The Protest Psychosis: How Schizophrenia Became a Black Disease*. Boston: Beacon Press.

Minear, M., & Park, D. C. 2004. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36, 630–633.

National Research Council. 2002. *Measuring Housing Discrimination in a National Study: Report of a Workshop*. Washington, DC: The National Academies Press.

———. 2010. *The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary*. Washington, DC: The National Academies Press.

Nguyen, Hannah-Hanh D., and Ann Marie Ryan. 2008. “Does Stereotype Threat Affect Test Performance of Minorities and Women? A Meta-Analysis of Experimental Evidence.” *Journal of Applied Psychology* 93 (6): 1314–34.

Orne, M. T. 2009. Demand characteristics and the concept of quasi-controls. *Artifacts in Behavioral Research* (ed. Robert Rosenthal and Ralph L. Rosnow., pp. 110–137). Oxford: Oxford University Press.

Osborne, Jason W. 2006. “Gender, Stereotype Threat, and Anxiety: Psychophysiological and Cognitive Evidence.” *Electronic Journal of Research in Educational Psychology* 4 (8): 109–38.

Peffley, Mark, Jon Hurwitz, and Paul M. Sniderman. 1997. "Racial Stereotypes and Whites' Political Views of Blacks in the Context of Welfare and Crime." *American Journal of Political Science* 41 (1): 30–60.

Pettigrew, Thomas F. 1979. "The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice." *Personality and Social Psychology Bulletin* 5 (4): 461–76.

Porter, Theodore. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. 1st. pbk. ed. Princeton N.J.: Princeton University Press.

Ross, Dorothy. 1992. *The Origins of American Social Science*. Ideas in Context. Cambridge: Cambridge University Press.

Rothschild, Jeffrey M, Christopher P Landrigan, John W Cronin, Rainu Kaushal, Steven W Lockley, Elisabeth Burdick, Peter H Stone, et al. 2005. "The Critical Care Safety Study: The Incidence and Nature of Adverse Events and Serious Medical Errors in Intensive Care." *Critical Care Medicine* 33 (8): 1694–1700.

Sanders, L. May 29, 2005. Diagnosis: Facial tic, weight loss, delusion. *The New York Times*. New York, NY.

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.

Selten, Jean-Paul, and Elizabeth Cantor-Graae. 2005. "Social Defeat: Risk Factor for Schizophrenia?" *The British Journal of Psychiatry* 187: 101–2.

Shojania, K. G. 2003. "Changes in Rates of Autopsy-Detected Diagnostic Errors Over Time: A Systematic Review." *JAMA: The Journal of the American Medical Association* 289 (21): 2849–56.

Sidanius, Jim, and Felicia Pratto. 2001. Social Dominance an Intergroup Theory of Social Hierarchy and Oppression. Cambridge: Cambridge University Press.

Steele, Claude M., and Joshua Aronson. 1995. "Stereotype Threat and the Intellectual Test Performance of African Americans." *Journal of Personality and Social Psychology* 69 (5): 797–811.

Steele, Claude M., Steven J. Spencer, and Joshua Aronson. 2002. "Contending with Group Image: The Psychology of Stereotype and Social Identity Threat." *Advances in Experimental Social Psychology* 34: 379–440.

Strakowski, S M, M Flaum, X Amador, H S Bracha, A K Pandurangi, D Robinson, and M Tohen. 1996. "Racial Differences in the Diagnosis of Psychosis." *Schizophrenia Research* 21 (2): 117–24.

Trierweiler, S J, H W Neighbors, C Munday, E E Thompson, V J Binion, and J P Gomez. 2000. "Clinician Attributions Associated with the Diagnosis of Schizophrenia in African American and Non-African American Patients." *Journal of Consulting and Clinical Psychology* 68 (1): 171–75.

Williams, David R., and Selina A. Mohammed. 2008. "Discrimination and Racial Disparities in Health: Evidence and Needed Research." *Journal of Behavioral Medicine* 32 (1): 20–47.