Commentary: What's in a Name? Statistical Discrimination and Disparities Research

Timothy P. Hofer

In this issue of *Health Services Research*, Maserejian, Lutfey, and McKinlay (2009) report on an experiment examining the reasons for gender differences in diagnostic decision making for coronary heart disease (CHD). The experiment utilizes actors of different genders speaking from identical scripts of a patient presenting with fairly classic CHD symptoms. The authors argue that the experimental design standardized the presentation and history available to providers, leaving two potential explanations for residual variability in diagnosis of CHD: differences due to different gender estimates of disease prevalence (a prior probability), and differences due to personal stereotypes and biases. To distinguish between these two explanations, they set out to determine whether physician estimates of prior probability of CAD explain differences in CAD diagnosis by gender.

This type of experimental approach is very appealing, as it allows for the manipulation of several key variables of interest while removing considerable confounding due to the differing content and form of symptom communication and miscommunication found in real life. Such complexities may be due to any number of social, cultural, and disease severity factors, not to mention myriad reasons specific to the particular interaction between a provider and patient. Given the ability to control for key confounders, well-designed vignette-based experiments are particularly interesting and thought provoking, although sometimes challenging to analyze and interpret.

Consistent with several prior studies cited in the article, the authors find that physicians reported a higher diagnostic certainty of the CHD diagnosis in men than in women performing the same vignette. However, including physician-specific estimates of the prevalence of CHD for men and women as main effects in the regression did not diminish the effect of gender. Given the standardized presentation of symptoms, and the lack of diminution of the

gender effect on certainty of CHD when controlling for prevalence, the authors conclude that “the influence of stereotypes and prejudices should be examined in future work.”

The paper raises several interesting issues that I will discuss ranging from how the CHD prevalence estimates are elicited, what they measure, and how they are analyzed to the general concept of “statistical discrimination” that is used to frame and interpret the experiment.

As acknowledged by the authors, one of the potential limitations of the design was the lack of nuance in the elicitation of the physicians’ gender-specific CHD prevalence estimates. For each physician, the researchers assessed only a relative estimate of whether CHD prevalence in men was higher, lower, or the same as in women. They argued that the use of absolute estimates would have been less reliable. A second possible issue is that CHD is increasingly recognized as a heterogeneous disease, and while the form manifested by plaque rupture, thrombosis, and myocardial infarction is much more common in males, the frequent presentation of CHD as angina is actually more common in women (Hemingway et al. 2008). Lack of specificity in the scenario eliciting the prevalence estimates could cause a loss of precision and accuracy in the measure of the physicians’ beliefs about the gender-specific prevalence of CHD.

Beyond the method of eliciting the prevalences, two very different measures of group prevalence have been used in the literature exploring how group prevalence estimates influence differences in health care. In a previous study (cited by the authors), the group prevalence estimates were derived from population health surveys and it was assumed that all physicians were equally aware of this externally generated prevalence measure (Balsa, McGuire, and Meredith 2005). In contrast, Maserejian, Lutfey, and McKinlay (2009) directly measure each individual physician’s estimate of the group prevalence, a measure which incorporates each physician’s knowledge about the statistic along with any prejudice or bias that may influence his or her estimate of the group prevalence. While using epidemiologic data is likely a more accurate measure of the true group prevalence, the latter measure is a better representation of the physicians’ beliefs about the group prevalence. Both the interpretation and desirability of using group prevalence estimates in clinical decision making would

Address correspondence to Timothy Hofer, M.D., VA Center for Practice Management and Outcomes Research, P.O. Box 13017, Ann Arbor, MI 48113-0170; e-mail: thofer@umich.edu.

seem to be quite different depending on how much the physician measure differs from the one based on national epidemiological sources.

Using this more direct measurement of the physicians' beliefs about the group prevalence, Maserejian, Lutfey, and McKinlay (2009) found that the prevalence estimates did not account for the gender differences in the certainty of CHD diagnosis when included as a main effect in the regression. However, to truly test for the effect of group prevalence on diagnostic certainty, the authors need to test for a cross-level interaction between the gender coefficient (a patient-level variable) and the physician-specific estimate of CHD prevalence (a physician-level variable). In the design, physician observations are clustered within patient video vignette. The physician-level CHD group prevalence variable can only account for some of the gender effect if it is differentially distributed across patients, which given the random assignment of physicians to patients should not occur in this study. An interaction term could test if the patient-level coefficient for gender varies as a function of the physician-specific prevalence estimate.

Finally, the effect of group prevalence estimates on clinical decision making is interpreted as one component of a concept described as "statistical discrimination." As a relative outsider to this field I was unfamiliar with this term, which nevertheless appears to be in relatively common use across a variety of disciplines. However, after reading more about it, I am not convinced that the concept is sufficiently specific to make it useful as a way to investigate treatment disparities. The term communicates little about a particular mechanism of discrimination and, indeed, encompasses a variety of different mechanisms that could account for differences in treatment, some of which seem poorly labeled as discrimination.

The term "statistical discrimination" appears to have been bequeathed to us by economists as a way to try to explain wage differences between social groups based on race or gender (Phelps 1972). Relying on Bayesian Decision Theory, an employer's wage decision is assumed to be based on a measurement of individual ability combined with the employer's prior knowledge of the group mean ability, with both of these measures weighted based on the reliability of the measure of the individual's ability. The theory argues that this decision making strategy should maximize utility for a rational decision maker. The following model is implied:

$$\text{Wage} = R \times \text{ability}_{(\text{individual estimate})} + (1 - R) \times \text{ability}_{(\text{group estimate})}$$

where R is the reliability which is a function of the variances of the two ability estimates, the between patient variance divided by the total variance or the sum of the between and within patient variance. Thus, the model contains four total parameters.

For diagnostic decision making in health care, the model becomes:

$$\begin{aligned} \text{Likelihood of Diagnosis} = & R \\ & \times \text{probability based on symptoms}_{(\text{individual estimate})} \\ & + (1 - R) \times \text{prior prevalence}_{(\text{group estimate})} \end{aligned}$$

Statistical discrimination was originally proposed to result when the individual characteristics of the members of one group are measured less reliably than those of another group, perhaps due to cultural or communication barriers (Phelps 1972). For members of that group, relatively less weight is put on the individual measurement and more on the group mean (in essence, information previously known about the group carries more weight as a consequence of the lack of reliability of the measure of an individual's ability). So far so good. Then things get messy.

Even though statistical discrimination originally was defined only to refer to the uncertainty in the measurement of individual ability, it was quickly generalized to include mechanisms that involve the other components of the Bayesian decision model. For example, gender differences in life insurance premiums do not involve differential uncertainty about individual measurement of risk but are solely based on differences in the group means represented by gender-specific group-based life expectancies. This specific example is presented as a form of statistical discrimination in which "discrimination is the agents' efficient response to asymmetric beliefs" about group characteristics (Moro 2009). Other developments of the concept invoke additional mechanisms such as risk aversion on the part of the employers as a reason why they might require a higher assessment of individual ability for the same wage from members of a group whose individual characteristics they perceive less precisely (Aigner and Cain 1977). Even more complicated dynamic models of statistical discrimination have been proposed involving feedback loops into several parts of the Bayesian decision model (Oettinger 1996).

In the health care disparities literature, the Institute of Medicine (IOM) report "Differences, disparities and discrimination" does not mention the term "statistical discrimination." It does, however, describe a form of discrimination corresponding to Phelps' original definition of statistical discrimination based on the differential reliability of measurement (referred to as "uncertainty" in

the report) (Smedley, Stith, and Nelson 2003). The report provides an example of differential treatment between a white and Latino patient where the mechanism is limited to uncertainty in the measurement of individual symptoms with constant group mean differences. By comparison, in the Maserejian, Lutfey, and McKinlay (2009) paper in this issue (as well as in the cited Balsa, McGuire, and Meredith 2005 paper) statistical discrimination is explicitly defined to include differences that arise solely from group mean differences (called the prevalence hypothesis) holding individual symptoms constant. The IOM report does not mention a prevalence hypothesis.

When one considers that there are only four parameters in the Bayesian Decision Theory model (the mean and variance of the measurement of individual characteristics along with the mean and variance of the group characteristic), the above multiplicity of mechanisms ascribed to the concept of statistical discrimination can be seen as potentially relating to all of these components. Since the term does not appear to describe a meaningful subset of mechanisms, it would seem more direct to just describe specific mechanisms that can have an impact on each of the components of the Bayesian Decision Theory model.

Apart from the lack of specificity in the use of the term, the words themselves making up the term seem problematic. The label "statistical" is in comparison to what? The entire framework, the Bayesian Decision Theory model, is a statistical model, and mechanisms relating to any of its components are statistical. More importantly, the use of the word "discrimination" is too broad. It includes mechanisms underlying treatment differences that are due to prejudice, bias, and stereotyping as well as mechanisms that produce treatment differences by maximizing utility when based on high-quality measurement and prior information. I do not see an advantage in using a term like "discrimination," which has a firmly negative connotation, to lump biases that lead to injustice in with appropriate behaviors that maximize patient utility. And by Bayesian Decision Theory, patient utility is maximized when individual and prior group information is correctly weighted in the face of measurement that is of high quality but inevitably less than perfectly precise. The danger of painting with such broad strokes can be seen in the medical errors/patient safety field, where the term "error," with its intuitive connotation of blameworthiness, was co-opted to include a broad range of "actions not completed according to plan." The term lost much of its meaning; rare blameworthy errors were used to garner attention and resources and the general stigmatization of all actions included in the broadened term has undermined efforts to facilitate

the transparency and openness necessary to solve the much more common complex system failures (Hofer 2000).

I would argue that a better way to define discrimination in the context of the Bayesian Decision Theory framework and the IOM disparities model would be to specify that disparities in clinical decision making arise only when decisions are not based on high-quality measurement and information. By high quality I mean measurement that is optimally accurate and precise. Within that category, discrimination occurs when the mismeasurement of either individual or group characteristics of members of a particular group is caused by ignorance, miscommunication, or bias.

Maserejian, Lutfey, and McKinlay (2009) have interestingly demonstrated considerable heterogeneity in physician estimates of CHD prevalence and some evidence of inconsistency in their use of prior information. Perhaps one of the most important points raised is whether the use of faulty knowledge of priors or faulty combination of prior prevalence estimates and individual symptoms could outweigh the utility-maximizing advantage of using prior information at all (Cahan et al. 2003). At the very least it provides an argument for increased clinical decision support in order to ensure the appropriate use of prior group-based prevalence estimates and their combination with individual characteristics measured at the clinical encounter. This seems the best way to standardize some of the inputs into clinical diagnostic decision making so as to prevent lack of knowledge, unwarranted stereotypes, and prejudice from contributing further to disparities in treatment.

REFERENCES

- Aigner, D. J., and G. G. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review* 30 (2): 175–87.
- Balsa, A. I., T. G. McGuire, and L. S. Meredith. 2005. "Testing for Statistical Discrimination in Health Care." *Health Services Research* 40 (1): 227–52.
- Cahan, A., D. Gilon, O. Manor, and O. Paltiel. 2003. "Probabilistic Reasoning and Clinical Decision-Making: Do Doctors Overestimate Diagnostic Probabilities?" *Quarterly Journal of Medicine* 96 (10): 763–9.
- Hemingway, H., C. Langenberg, J. Damant, C. Frost, K. Pyorala, and E. Barrett-Connor. 2008. "Prevalence of Angina in Women versus Men: A Systematic Review and Meta-Analysis of International Variations across 31 Countries." *Circulation* 117 (12): 1526–36.
- Hofer, T. P., E. A. Kerr, and R. A. Hayward. 2000. "What Is an Error?" *Effective Clinical Practice* 3 (6): 261–9.

- Maserejian, N. N., K. E. Lutfey, and J. B. McKinlay. 2009. "Do Physicians Attend to Base Rates? Prevalence Data and Statistical Discrimination in the Diagnosis of Coronary Heart Disease." *Health Services Research*, doi: 10.1111/j.1475-6773.2009.01022.x.
- Moro, A. 2009. "Statistical Discrimination." In *The New Palgrave Dictionary of Economics*, online edition, edited by S. N. Durlauf and L. E. Blume. New York: Palgrave, Macmillan [accessed on August 03, 2009]. Available at http://www.dictionaryofeconomics.com/article?id=pde2009_S000544
- Oettinger, G. S. 1996. "Statistical Discrimination and the Early Career Evolution of the Black-White Wage Gap." *Journal of Labor Economics* 14 (1): 52–78.
- Phelps, E. S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659–61.
- Smedley, B. D., A. Y. Stith, and A. R. Nelson. 2003. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Washington, DC: National Academy Press.