American Journal of Law & Medicine, 43 (2017): 193-208
American Society of Law & Medical Ethics, © 2017 The Author(s)
Boston University School of Law
DOI: 10.1177/0098858817723659

STATISTICS, ADJUSTED STATISTICS, AND MALADJUSTED STATISTICS

Jay S. Kaufman[†]

Statistical adjustment is a ubiquitous practice in all quantitative fields that is meant to correct for improprieties or limitations in observed data, to remove the influence of nuisance variables or to turn observed correlations into causal inferences. These adjustments proceed by reporting not what was observed in the real world, but instead modeling what would have been observed in an imaginary world in which specific nuisances and improprieties are absent. These techniques are powerful and useful inferential tools, but their application can be hazardous or deleterious if consumers of the adjusted results mistake the imaginary world of models for the real world of data. Adjustments require decisions about which factors are of primary interest and which are imagined away, and yet many adjusted results are presented without any explanation or justification for these decisions. Adjustments can be harmful if poorly motivated, and are frequently misinterpreted in the media's reporting of scientific studies. Adjustment procedures have become so routinized that many scientists and readers lose the habit of relating the reported findings back to the real world in which we live.

I. STATISTICS

Quantitative summaries are essential for understanding our world, learning about causes and consequences of human actions and policies, and assessing equity and justice.[1] For this reason, there is a robust tradition across all complex societies of regularly collecting census data, vital statistics, and other observations to facilitate surveillance of the population and monitor trends over time.[2] Quantitative measures also play an essential role in biomedical and social science research in order to

[†]Professor and Canada Research Chair in Health Disparities, Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada. Supported by the Canada Research Chairs Program. Drs. Osagie Obasogie, Hailey Banack, Nicholas King, and Joanna-Trees Merckx provided generous and insightful comments on an earlier draft of the manuscript.

[1] *See* THE MUTUAL CONSTRUCTION OF STATISTICS AND SOCIETY 66 (Ann Rudinow Saetnan, Heidi Mork Lomell & Svein Hammer eds., 2010) [hereinafter MUTUAL CONSTRUCTION OF STATISTICS] (treating statistical knowledge as producing the "spatiotemporal framework for the experience of populations, nations, classes, and social problems" and finding that "statistics do the work of *holding together* knowledge, practice, and the State.") (emphasis in original).

[2] *See generally* ALAIN DESROSIÈRES, THE POLITICS OF LARGE NUMBERS: A HISTORY OF STATISTICAL REASONING (Camille Naish trans., 1998) (examining the long history of statistics and its connection with the construction, unification and administration of the State).

estimate consequences of human decisions, such as prescribing a medicine or enacting a law.[3] Without regular and reliable measures of this kind, we would be blind to the prevailing conditions of the population, and unable to make rational decisions to improve the length and quality of life for members of the society.[4]

Statistics serves a fundamental role in this endeavor, providing a science for understanding the relation between samples and the populations from which they derive, and for describing variability in observations and modeling this variability to understand aspects of structure, mechanism, causal effect and attribution. Such inferences are essential not only for understanding our world as it exists, but also for making predictions about the future. They are essential for adjudicating disputes that rest on issues of distributive justice, since attribution is a function of causal theories.[5]

A basic activity of statisticians and their related professions (e.g. epidemiologists, demographers, and quantitative social scientists) is to relate observations from random samples to the unobserved quantities in the total population, and to build models that relate measures to one another in the form of associations.[6] We estimate parameters that have a magnitude, such as the average measurement in a group, or the correlation between two factors. This is generally some function of the individual data points that summarizes information into a single value that represents the aggregate quantity of interest. There is also a variance, which represents something about the distribution of data points across the sample or across the population.[7] When a random sample of values is observed and we wish to make inferences about the value of a quantity of interest in the whole population, the variance of the summary across repeated samples tells us something about how much uncertainty is attributable to sample selection.[8] For example, if we poll ten randomly selected citizens about their support for a political party, we should know less securely the true level of population support than if we randomly poll 100 citizens. This additional uncertainly due to the smaller sample is revealed by a larger variance of the observed proportion over repeated samples.

A key conceptual foundation of statistics is the notion of the distribution, which is a mathematical depiction of the frequency of observations.[9] Recognizing phenomena and processes that follow defined distributions is a routine aspect of statistical inference, since the distribution has a precise mathematical expression as a function of specified parameters.[10] For example, coin flips follow a "binomial distribution," controlled by a parameter referring to the probability of observing a "heads," as well as a number of times the coin is flipped. From these two numbers we can define the variance of the count of heads, or the variance around an estimate of a population

---

[3] MUTUAL CONSTRUCTION OF STATISTICS, *supra* note 1, at 66 ("Statistics has become, at least in some forms of practice, the epistemic flagship of the modern sciences – be it in biology, physics, informatics, or sociology.").

[4] *See generally* STATISTICS IN SOCIETY: THE ARITHMETIC OF POLITICS (Daniel Dorling & Stephen Simpson eds., 1999) (explaining the need for widespread comprehension of statistical insights and skills on topics such as gender, ethnicity, religion, poverty, race, health, education, unemployment and politics, among others).

[5] *See generally* Sander Greenland, *Concepts and Pitfalls in Measuring and Interpreting Attributable Fractions, Prevented Fractions, and Causation Probabilities*, 25 ANNALS EPIDEMIOLOGY 155 (2015) (explaining how attributive theory is intimately connected to the function and proliferation of causal theories and models).

[6] *See* SUSAN DEAN & BARBARA ILLOWSKY, INTRODUCTORY STATISTICS 7 (2017) (ebook).

[7] *Id*. at 113.

[8] *Id*. at 116.

[9] *Id*. at 105–106.

[10] *Id*. at 240 ("A discrete probability distribution function has two characteristics: 1. Each probability is between zero and one, inclusive. 2. The sum of the probabilities is one.").

proportion based on a small sample. Once a distribution has been defined as a process that gave rise to the data observed, then the individual observations can be smoothed into summaries that have defined expectations and variances, and the relations between quantities will also have defined characteristics.[11] Models for the relations between quantities can then become increasingly elaborated with assumptions, such as linearity or independence, where these are convenient for smoothing over noisy patterns, or extrapolating over regions with few observations.[12]

II. ADJUSTED STATISTICS

When samples are randomly selected, the resulting summaries and associations refer to the real world of the population from which the samples were drawn. Often, however, we *adjust* these estimates in some way, perturbing the estimated value in a specific way to make it more relevant or useful in some sense. Such adjustments are a major occupation within the science of statistics, with the simple depiction of the "raw" or "crude" values considered to be rather trivial.[13] Browsing through the journals of statistics, epidemiology, or related fields, one will find that most of the techniques described are new or improved methods for making such adjustments.[14] If we make an analogy between statistics and cooking, then obtaining the crude values is something like grocery shopping, a necessary process, but not the least bit glamorous. The real art and marvel of the craft of cooking, however, is expertly combining and modifying the ingredients, drawing out flavors, and modifying textures. This is adjustment.

Adjustments can be performed to address some impropriety in the sampling or measurement, toward a more accurate portrayal of the real world population from which the samples were drawn.[15] This is the case, for example, with adjustments to the census to address population undercounts,[16] with adjustments that are made to correct for measurement error in nutritional surveys,[17] or for the impact of missing data due to respondents who are lost during follow-up.[18] More commonly, however, the adjustment conducted by the statistician is not to improve the depiction of the real world, but rather to depict a quantity in an imaginary world.[19] It is the crude observation that reflects the events that occurred in the real world, but this number may be deemed less relevant to some hypothetical question, therefore justifying the adjustment.

---

[11] Sander Greenland, *Summarization, Smoothing, and Inference in Epidemiologic Analysis*, 21 SCANDINAVIAN J. SOC. MED. 227, 228 (1993) ("Smoothing is usually viewed as the combination of data with a model to obtain data expected under the model – more precisely, the data one should expect to see in a replicate of the study if the model is correct and the replication is perfect with respect to all identified study and subject characteristics.").

[12] *Id*. at 230.

[13] *See* Niels Keiding & David Clayton, *Standardization and Control for Confounding in Observational Studies: A Historical Perspective*, 29 STAT. SCI. 529 (2014).

[14] *See id*. (describing the emergence of the regression modeling approach and the refinement of the weighting approach for confounder control during the twentieth-century).

[15] *Id*. at 529 (explaining that methods of standardization of rates compare predicted marginal summaries to target populations).

[16] THE 2000 CENSUS: COUNTING UNDER ADVERSITY (Constance F. Citro, Daniel L. Cork & Janet L. Norwood eds., 2004).

[17] Laurence S. Freedman et al., *Dealing with Dietary Measurement Error in Nutritional Cohort Studies*, 103 J. NAT'L CANCER INST. 1086, 1089–90 (2011).

[18] Jennifer Weuve et al., *Accounting for Bias Due to Selective Attrition: The Example of Smoking and Cognitive Decline*, 23 EPIDEMIOLOGY 119, 121 (2012).

[19] *See* Kieding & Clayton, *supra* note 13, at 542.

In fact, while routine statistical surveillance such as reported by government agencies may be presented as crude frequencies in order to best depict the real world, all etiologic or causal estimates are based on hypothetical conditions, since the question is inherently a "what if" construction rather than a "what is" construction.[20] For example, the question of whether smoking causes lung cancer is essentially a contrast between the risk of lung cancer for an individual if she smokes compared to the risk for that same individual if she doesn't smoke. As such, one half of this contrast is always unobserved, a fact that has been dubbed "[t]he fundamental problem of causal inference . . . ."[21] At the aggregate level, if we observe a sample that is a mix of smokers and non-smokers, and we ask what would be the lung cancer rate if they had all smoked versus if none of them had smoked, then both parts of this contrast are unobserved.[22]

Even for surveillance, it is often the case that the hypothetical world is more interesting or relevant in some way than the factual world. The most common example of this is age standardization, which is a routine adjustment in nearly all vital statistics reporting, despite the fact that it refers to imaginary populations when the presumed goal is depiction of actual event frequencies.[23] Consider the following example from Schoenbach and Rosamond, shown in Appendix Table 1, comparing deaths per 1,000 white women per year in Miami, Alaska, and the United States.[24] The true mortality observed in Miami was 8.92 deaths per 1,000 in that year, compared to an observed rate of 2.67 deaths per 1,000 in Alaska.[25] The United States as a whole experienced 8.13 deaths per 1,000.[26] Barring any misclassifications of vital status, race or gender, these numbers represent the true summary, but in many ways not the most interesting one.

The nuisance in this case is the age structure. From the summary numbers, it appears to be much safer to live in Alaska, where only a third as many white women die each year, but this is an artifact of a very different age structure in the two populations, also depicted in Appendix Table 1.[27] Indeed, by looking at the death rate within each age group, it can be observed immediately that the mortality rates are approximately the same between Miami and Alaska.[28] The overall death rate in Miami is higher simply because there is a greater proportion of old people in Miami and old people die at a higher rate.[29] The judgment that the crude rate is an inadequate summary, therefore rests on the notion that Miami should not be punished in its assessment simply for having more elderly residents, since the age-stratified mortality is comparable. The value judgment underlying this discomfort is that age is a permissible cause of mortality, and therefore should be extracted from the summary so we don't mix the effect of age with the effect of location. A key insight is that all statistical adjustments require a value judgment concerning differences that are

---

[20] *See id.* at 541.

[21] GUIDO W. IMBENS & DONALD B. RUBIN, CAUSAL INFERENCE IN STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES 24 (2015).

[22] George Maldonado & Sander Greenland, *Estimating Causal Effects*, 31 INT'L J. EPIDEMIOLOGY 422, 424, 428 (2002).

[23] Richard J. Klein & Charlotte A. Schoenborn, *Age Adjustment Using the 2000 Projected U.S. Population*, 20 HEALTH PEOPLE 2010 STAT. NOTES 1, 1 (2001).

[24] VICTOR J. SCHOENBACH & WAYNE D. ROSAMUND, UNDERSTANDING THE FUNDAMENTALS OF EPIDEMIOLOGY: AN EVOLVING TEXT 132 (2000) (ebook).

[25] *Id.*

[26] *Id.*

[27] *Id.*

[28] *Id.*

[29] *Id.*

permissible and those that are impermissible, in order to eliminate the nuisance of the first category and focus on the disparity due to the second category.[30]

The traditional solution in this case is age-standardization, in which a new summary is constructed as a weighted average of the stratum-specific rates, with weights taken from a "standard" reference population, such as the entire United States in 1970.[31] We apply the same set of weights to the age-specific rates of Alaska and Miami so that the summary (age-adjusted) death rate will be independent of differences in the age distribution of the two populations. This is arguably a more interesting and relevant comparison, but it must be remembered that this adjusted summary estimate is no longer an observation about the real world. Rather, the adjusted rates represent the crude death rates that Miami and Alaska *would have experienced* if they had both had (counter to fact) the same age distribution as the 1970 U.S. white female population. The new adjusted rate is a fiction, but a convenient one because it removes the distorting effect of a nuisance covariate to allow a "fairer" comparison of the death rates. Applying the formula above, we obtain age-standardized death rates for Miami equal to 6.92 per 1,000, and for Alaska equal to 6.71 per 1,000. These are approximately the same, and reveal that one location is not more dangerous than the other, and in fact that both match very closely to the United States as a whole. In this way the adjusted contrast reveals that although many more white women died in Miami than in Alaska, a more informative comparison of these rates shows that at any given age, risk is not elevated importantly in either location.

The ubiquitous application of such adjustments should not obscure the fact that the adjusted rates are imaginary. They can be easily changed by altering something as arbitrary as the standard population that is chosen to provide the weights. An impressive example of this phenomenon is provided by Krieger and Williams.[32] After decades of using the 1970 census as the standard population in U.S. official statistics, it was President George W. Bush's administration that oversaw the switch to the new 2000 standard after the census of that year became available.[33] Because the U.S. population in 2000 was substantially older than it was in 1970, older age strata were subsequently given more weight in the standardization formula.[34] For somewhat artifactual reasons, racial, and socioeconomic disparities tend to be more modest in older populations when measured using ratios.[35] Therefore, the result of the switch in the reference populations to an older standard was that it reduced all reported disparities overnight.[36] This resulted in the Bush administration reporting an immediate reduction in disparities under their watch, even though this arose from a statistical procedure rather than any actual improvement for any individual.[37]

---

[30] Sam Harper et al., *Implicit Value Judgments in the Measurement of Health Inequalities*, 88 MILBANK Q. 4 (2010) (discussing the involvement of value judgments in the adjustment of statistical analyses).

[31] $Standardized\ Rate = \frac{(r_1N_1+r_2N_2+r_3N_3+\ldots+r_nN_n)}{(N_1+N_2+N_3+\ldots+N_n)}$, where r_k is rate in the k^{th} stratum of the study population and N_k is the number of people in the k^{th} stratum of the standard population.

[32] Nancy Krieger & David R. Williams, *Changing to the 2000 Standard Million: Are Declining Racial/Ethnic and Socioeconomic Inequalities in Health Real Progress or Statistical Illusion?,* 91 AM. J. PUB. HEALTH 1209 (2001).

[33] *See* Klein & Schoenborn, *supra* note 23, at 1–2.

[34] *See* Krieger & Williams, *supra* note 32, at 1211.

[35] Jay S. Kaufman et al., *The Relation Between Income and Mortality in U.S. Blacks and Whites*, 9 EPIDEMIOLOGY 147, 148 (1998).

[36] *Id.* at 152.

[37] *But see Age Standardization of Death Rates: Implementation of the Year 2000 Standard*, NAT'L VITAL STAT. REP. (CDC), Oct. 7, 1998 (explaining the impact of the implementation of the year 2000 population standard on statistical variability).

III. PERMISSIBLE AND IMPERMISSIBLE FACTORS

The normative value that older people can be expected to die more frequently than younger people seems difficult to contest, but agreeing on a list of permissible variables quickly becomes more difficult once we go beyond age and sex. As an example, consider an analysis of the National Assessment of Educational Progress (NAEP), published by the Urban Institute in 2015.[38] The NAEP is a standardized test regularly administered to a nationally-representative sample of U.S. students, and there are crude average test scores available for each state.[39] The primary interest of the Urban Institute, however, is on the role of state educational policy in making improvements to student performance.[40] They argue that this assessment of state policy is not directly-observable from the crude test scores because of demographic differences between states, such as the proportion of students in poverty or the proportion that are black.[41]

The Urban Institute therefore calculated adjusted NAEP scores, based on the 2013 results, to account for these differences in ostensibly permissible factors.[42] Louisiana, for example, actually performed 47th out of 50, but its adjusted rank was 27th.[43] Likewise, Texas ranked 3rd in the nation after adjustment, despite a true performance that ranked 32nd.[44] In the adjusted world, Texas and Louisiana vastly outperformed California and Michigan, with the interpretation that this analysis makes more relevant comparisons that are balanced on race and poverty in a way that the real world is not.[45] Just as Miami was not to be punished in the previous example for having more elderly residents, the same logic suggests that Texas should not be punished for having more Hispanics, or Louisiana for having more blacks.[46] From the perspective of evaluating the success or failure of a state's education policy, this seems sensible. Each state must educate the students that reside there, whether poor or minority or whatever other characteristics they have. But when one considers Louisiana a success, accounting for their demographic challenges, despite being in objective terms at the very bottom, one risks normalizing the expectation that poor and minority kids should perform poorly. If these differences are accepted as permissible, it removes those factors from critical scrutiny, conditioning away any inequality attributable to them as a given, or as a mere nuisance or artifact. Like the photographer's adjustment that can hide an ugly blemish, statistical adjustment, too, can sweep many unpleasant realities under the carpet where they are no longer seen.

While statistical adjustments in the context of disparities research are notoriously problematic because of controversies over which covariates are deemed permissible

---

[38] Matthew M. Chingos, Urban Inst., Breaking the Curve: Promises and Pitfalls of Using NAEP Data to Assess the State Role in Student Achievement (2015), http://www.urban.org/sites/default/files/publication/72411/2000484-Breaking-the-Curve-Promises-and-Pitfalls-of-Using-NAEP-Data-to-Assess-the-State-Role-in-Student-Achievement.pdf [https://perma.cc/5E6U-XWNH].

[39] *NAEP Overview*, NAT'L CTR. FOR EDUC. STAT., https://nces.ed.gov/nationsreportcard/about/ [https://perma.cc/LXU6-SX7B] (last updated Mar. 30, 2016).

[40] CHINGOS, *supra* note 38, at 1.

[41] *Id.* at 3. Note that the label "blacks" is used in this paper to refer to African Americans, and "whites" to European Americans. Where some of the papers referenced have used other terminology, I have generally converted the terminology to these labels for reasons of simplicity and consistency. In all cases, the terms refer to self-identified groups within the United States.

[42] *Id*. at 4.

[43] *Id*. at app. A, Table A.1.

[44] *Id*.

[45] *Id*. at 2.

[46] *Id*. at 4–5.

and how to interpret adjusted estimates,[47] there has been remarkably little formalization of this problem in the statistics literature itself. A notable exception is a rather remarkable paper by Duan et al. in which this problem is delineated and explored, and which shows the paradigm to be considerably more challenging than routine application would suggest.[48] Focusing on disparities in the provision and receipt of clinical services, the authors cite an Institute of Medicine (IOM) report which defines such disparities as differences in the quality of health care that are not due to "access-related factors or clinical needs, preference, and appropriateness of intervention."[49] The intent here is obviously to remove from consideration any factors that are outside the control of the healthcare provider. The authors note that the definition is explicitly causal, by requiring attention to what is due and what is not due to various factors.[50] It is also explicitly value-laden, because it considers as relevant for policy intervention only differences that are due to some kind of inappropriate or unjustifiable discrimination on the part of the clinician.[51] That different populations might have varying access to care or unequal needs is off the table for the purposes of this question.

Duan and colleagues then propose a statistical framework in which the covariates of interest have been classified into "allowable" and "non-allowable" categories with respect to a specific research goal.[52] Allowable covariates are considered to be a justifiable cause of difference and hence should be statistically adjusted away before reporting a disparity.[53] One then only reports a disparity due to the remaining non-allowable covariates, a disparity which is different from what exists in the real world, but is what would be observed in an imaginary world in which all allowable factors were balanced between groups.[54] For example, if an outcome is affected by tastes and preferences, it is allowable, so that a person who is obese simply because he enjoys eating is not considered disparate. If an outcome is affected by constraints and barriers that defy the will of the individual, however, then it is not allowable, such as a person who is obese because healthy food is not readily available in his neighborhood. Obviously there are many practical problems in defining this demarcation, not least of which is the need to know the causes of this individual's obesity. Then there is the values-laden task of separating these causes into those that are permissible, and those that represent some moral affront. Whether such a consensus around this demarcation can actually be achieved is unclear, and even more worrisome is that these justifications are seldom described explicitly in an analysis. Most often, authors simply provide a list of control variables with no justification whatsoever.[55]

---

[49] *See generally* Jay S. Kaufman et al., *Socioeconomic Status and Health in Blacks and Whites: The Problem of Residual Confounding and the Resiliency of Race*, 8 EPIDEMIOLOGY 621 (1997) (discussing potential sources for residual confounding in statistical adjustments based on socioeconomic factors); Tyler J. VanderWeele & Whitney R. Robinson, *On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables*, 25 EPIDEMIOLOGY 473 (2014) (analyzing the effect of adjustment to socioeconomic distributions when race is used as an exposure variable).

[48] Naihua Duan et al., *Disparities in Defining Disparities: Statistical Conceptual Frameworks*, 27 STAT. MED. 3941 (2008).

[49] *Id.* at 3942 (quoting Inst. of Med., Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care 32 (Brian D. Smedley et al. eds., 2002)).

[50] *Id*.

[51] *Id*.

[52] *Id.*

[53] *See id.*

[54] *See id.*

[55] *See, e.g.*, Greenland, *supra* note 5 (using examples of control group patients without providing justification); Keiding & Clayton, *supra* note 13 (controlling for variables in order to make comparisons, but failing to provide justification for comparison); Eric J. Tchetgen Tchetgen, *Identification and Estimation of*

Even more troublesome is that this distinction should ideally vary from study to study, depending on the specific scientific goals. So for example, although the Institute of Medicine considered inequality in access to care to be permissible in their investigation of disparities at the point of care, inequality in access to care should not be permissible to an investigator of health services systems.[56] The distinction is rooted in some notion of *blame*, which is extra-scientific, i.e. a moral or ethical judgment based on values. Statistical adjustment is simply the tool by which we avoid blaming individuals or institutions for factors that are understood to be in some way outside of their control or irrelevant to narrow considerations of justice and equity. As such, it is rather remarkable that these decisions go almost completely unexamined in biomedical and social science publications.

A more profound technical concern raised by Duan and colleagues is that since the adjustment is based on an explicitly causal model, it must also respect the causal order of the variables, because if X causes Y, then one cannot logically condition on a fixed value of Y while varying X.[57] For example, if we ask what is the effect of lifelong heavy smoking versus lifelong non-smoking only among ninety year old men, the question is impossible to answer in any simple way, because there will be fewer ninety year old men if the group is assigned to smoke instead of not smoke.[58] In this way, adjustments not only create an imaginary population, but potentially one that is not even plausibly observed if it defies the natural relations between factors. Thus for allowable variables X_A and non-allowable variables X_N, a major point of the Duan et al paper is to show that modeled disparities can be very different if we consider that X_A can affect X_N (in which case a conditional disparity must be modeled) or if X_N can affect X_A (in which case a marginal disparity must be modeled).[59] The appropriate formula for the marginal disparity is actually rather complicated, and not widely known. Therefore, most researchers are implicitly assuming X_A causes X_N, but probably without being aware of the distinction. This may often be true, but certainly not always. For example, it may be that access to care (X_A) affects blood pressure (X_N), but if high blood pressure leads to stroke, then this may in turn hinder access to care in the future. In this case, the adjusted disparity reported is not the one that would really be observed under a true intervention to eliminate non-allowed inequalities. Duan et al give worked examples in which the adjusted disparity changes direction based on the marginal versus conditional distinction.[60] This phenomenon is related to "Simpson's Paradox,"[61] and there are many real-world examples of this in the literature.[62]

Survivor Average Causal Effects, 33 STAT. MED. 3601 (2014) (utilizing control variables to account for confounding problems without providing rationale behind controls).

[56] INST. OF MED., UNEQUAL TREATMENT: CONFRONTING RACIAL AND ETHNIC DISPARITIES IN HEALTH CARE 77–79 (Brian D. Smedley et al. eds., 2002).

[57] *See generally* Enrique F. Schisterman et al., *Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies*, 20 EPIDEMIOLOGY 488 (2009) (utilizing a casual model analysis to illustrate and clarify the definition of overadjustment bias).

[58] Tchetgen Tchetgen, *supra* note 55, at 3602.

[59] *See* Duan et al., *supra* note 548, at 3944–55.

[60] *See generally* Duan et al., *supra* note 50.

[61] Miguel A. Hernán, David Clayton & Niels Keiding, *The Simpson's Paradox Unraveled*, 40 INT'L. J. EPIDEMIOLOGY 780 (2011).

[62] *See, e.g.*, Alexander Persoskie & Bryan Leyva, *Blacks Smoke Less (and More) than Whites: Simpson's Paradox in U.S. Smoking Rates, 2008 to 2012*, 26 J. HEALTH CARE POOR & UNDERSERVED 951, 952 (2015).

IV. MALADJUSTED STATISTICS

Two additional examples taken from the recent racial disparities literature help to illustrate these concerns as they arise in practice. The first is a new study by Beavis and colleagues on racial disparities in age-standardized cervical cancer mortality in the United States.[63] The authors observed that previously published surveillance of this inequality had not taken into consideration the race-specific prevalence of hysterectomy, a procedure which removes women from the population at risk.[64] The logic is rather unimpeachable: just as men are not included in the denominator because they lack a cervix, women who have had a hysterectomy are also not at risk for cervical cancer, and therefore should also be removed from the risk pool.[65] The authors therefore obtained data on the prevalence of hysterectomy for adult women from a nationally representative survey, stratified by age, state, year and race.[66] Because person-time not at risk was subtracted from the denominator of the rates, the resulting values were higher for both races after correction.[67] For black women, this changed the age-standardized rate from 5.7 to 10.1 per 100,000 per year, and for white women from 3.2 to 4.7 per 100,000 per year.[68] Thus, without the correction, the racial mortality disparity was underestimated by forty-four percent. Previously reported racial disparities for this outcome were severely underestimated, and the magnitude of this discrepancy also increased with age, because hysterectomy prevalence increases with age, and rises more steeply in black women than in white women.[69]

The analysis is expertly conducted and reported, but the authors never consider the ramifications of defining hysterectomy as a permissible covariate that should be adjusted away. Why should there be such a profound racial disparity in this procedure in the first place?[70] The published analysis shifts attention away from this aspect of the problem and focuses on the narrower question of what proportion of women with a cervix end up dying of cervical cancer.[71] This ignores the indisputable fact that hysterectomy does really prevent the outcome. The higher rate of hysterectomy in black women is not an artifact, but a real part of the complete picture of racial disparity. Taking this to the extreme to make the point, imagine that hysterectomy were performed on 100% of black women. This would drive the cervical cancer mortality rate in this group to zero, and the racial disparity would be reversed. Obviously this is not a desirable outcome, however, because this specific method of preventing the disease has other adverse consequences.[72] Nonetheless, the tangible implications of hysterectomy policy for cervical cancer and the clear racial disparity in

---

[63] Anna L. Beavis et al., *Hysterectomy-Corrected Cervical Cancer Mortality Rates Reveal a Larger Racial Disparity in the United States*, 123 CANCER 1044 (2017).

[64] *Id*. at 1044.

[65] *Id.*

[66] *Id* at 1045.

[67] *Id* at 1046.

[68] *Id.*

[69] *Id.* at 1047.

[70] *See* Katharine M. Esselen et al., *Health Care Disparities in Hysterectomy for Gynecologic Cancers*, 126 OBSTETRICS & GYNECOLOGY 1029 (2015) (concluding that there were striking racial disparities associated with the use of minimally invasive hysterectomy for uterine and cervical cancers).

[71] Beavis et al., *supra* note 65, at 1044.

[72] Jemma Mytton et al., *Removal of All Ovarian Tissue Versus Conserving Ovarian Tissue at Time of Hysterectomy in Premenopausal Patients with Benign Disease: Study Using Routine Data and Data Linkage*, 356 BRIT. MED. J. 1, 5 (2017).

both outcomes make it a real component of the overall disparity rather than a mere statistical nuisance.[73]

Confusion over this point is also evident in the media attention devoted to this article.[74] The press release carried in many online and print publications declared, "[A] new analysis reveals that for most women, the risk of dying from cervical cancer is higher than previously thought."[75] This is not factually correct, because the prevention of cervical cancer via hysterectomy is a characteristic of the real world, even if it is not a characteristic of the imaginary world depicted by the adjusted analysis.[76] The frequency of deaths is the central fact that is actually observed, and the absence of deaths in some sub-groups in the population is also truly observed.[77] These are both characteristics of the real world. Excluding them from the calculation creates an imaginary population that may be more relevant for some scientific or policy questions, but this distinction needs to be kept in mind to avoid confusing language like the assertion that "the risk of dying from cervical cancer is higher."[78] In fact, the apparent increase comes from simply removing some people from the denominator, not by observing any additional deaths in the real world.

The second example is a newly-published study of racial disparities in asthma by Nyenhuis and colleagues.[79] The authors noted that blacks have a greater burden from asthma compared with whites in the United States, and wondered whether the pattern of airway inflammation differs between these racial groups as some explanation for this disparity.[80] Therefore they compared the saliva and mucus coughed up by black and white subjects, stratified by inhaled corticosteroid use, looking for a difference in the proportion of certain white blood cells.[81] Specifically, they wanted to know if there was a difference in "eosinophilic inflammatory phenotype," defined as sputum eosinophil frequency above or below two percent.[82] Among 1,018 participants, black subjects (n=264) had worse lung function (80% vs. 85% FEV1 predicted) greater total IgE levels (197 vs. 120 IU/mL), and a greater proportion with uncontrolled asthma (43% vs. 28%) compared with white subjects (n=754).[83] Most subjects (922 out of 1,018) were using inhaled corticosteroids (95% of blacks and 89% of whites).[84]

The key finding in this study was that the type of airway inflammation was *not* significantly different between black and white subjects in either treatment group.[85] In those taking corticosteroids, the proportions with the eosinophilic type were 19% vs.

---

[73] Sarah M. Temkin et al., *The End of the Hysterectomy Epidemic and Endometrial Cancer Incidence: What Are the Unintended Consequences of Declining Hysterectomy Rates?*, 6 FRONTIERS ONCOLOGY 1, 3 (2016).

[74] *See, e.g.*, Wiley, *Cervical Cancer Mortality Rates May Be Underestimated*, SCIENCEDAILY (Jan. 23, 2017), www.sciencedaily.com/releases/2017/01/170123094748.htm [https://perma.cc/WW35-GRSB].

[75] *Id.*

[76] *See, e.g.*, Anne F. Rositch et al., *Increased Age and Race-Specific Incidence of Cervical Cancer After Correction for Hysterectomy Prevalence in the United States from 2000 to 2009*, 120 CANCER 2032, 2035 (2014).

[77] *Id.*

[78] Wiley, *supra* note 76.

[79] Sharmilee M. Nyenhuis et al., *Race is Associated with Differences in Airway Inflammation in Patients with Asthma*, J. ALLERGY CLINICAL IMMUNOLOGY, 1 (2017).

[80] *Id.* at 2.

[81] *Id.* at 4.

[82] *Id.* at 3.

[83] *Id.* at 3.

[84] *Id.* "FEV1" is the estimated volume of air that can be forced out of the lungs in one second, expressed as a percent of the total exhaled volume, and "IgE" is immunoglobulin E, the component of the immune system that drives allergic reactions.

[85] *Id.* at 5.

16% (p=0.28), and in those not taking corticosteroids, the proportions were 39% vs. 35% (p=0.65).[86] As the unadjusted observations, these proportions represent the state of the real world.[87] As typically happens, however, the authors were not as interested in the real world as they were in an adjusted world, and so they fit a logistic regression model to condition on measured covariates.[88] The selected covariates were age, sex, atopic status (positive skin test), body mass index (BMI in kg/m^2), lung function (predicted FEV1%), and degree of control of the disease (controlled vs. uncontrolled).[89] Conditional on all of these factors, black subjects had 58% higher odds of exhibiting eosinophilic airway inflammation than white subjects in the corticosteroid treated group (p =0.046) but not in the untreated group (p = 0.98).[90] The authors concluded that "African American subjects exhibit greater eosinophilic airway inflammation, which might explain the greater asthma burden in this population."[91]

Of course that last sentence is factually incorrect. In the real world, there was no difference whatsoever between race groups in the inflammatory phenotype.[92] It was only in an imaginary world constructed in a statistical model in which six additional factors were balanced between the race groups that the predicted distribution of this characteristic became differential.[93] The fact that the supposed racial difference only emerged in an adjusted model, not in the actual data, is completely omitted from the press release and extensive news coverage of the publication.[94] Indeed, the press release went even further, asserting that "African Americans may be less responsive to asthma treatment and more likely to die from the condition, in part, *because they have a unique type of airway inflammation* . . . ."[95] Even in the adjusted model, eosinophilic phenotype was by no means *unique* to black subjects, making this assertion an outrageous distortion of the published findings.[96]

We have already noted that the imaginary population of an adjusted estimate can be more relevant for some scientific or policy questions than the unadjusted population of the real world, but it is important to note that adjustments can also distort the association in a manner that make the results *less* relevant or interpretable.[97] Confounding adjustments work by balancing baseline variables across the exposure groups so that they become uncorrelated with the exposure, just as they would be in a randomized experiment.[98] To achieve this goal of covariate balance, however, it is essential to avoid adjusting for factors affected by the exposure.[99] The exposure of

---

[86] *Id.* at 3–4.

[87] *See id.* at 3–4.

[88] *Id.* at 405. Logistic regression is a generalized linear model with a logit link and a binomial error distribution, where logit refers to ln[p/(1-p)], p is the risk of the outcome, and p/(1/p) is referred to as the odds of the outcome. Exponentiated regression coefficients from this model have an odds ratio interpretation. *See, e.g.*, DAVID W. HOSMER, JR. ET AL., APPLIED LOGISTIC REGRESSION 1-33 (3d ed. 2013).

[89] Nyenhuis et al., *supra* note 79, at 3.

[90] *Id.* at 5, 7.

[91] *Id.* at 2.

[92] *See supra* text accompanying notes 85–87.

[93] *See supra* text accompanying note 89.

[94] Univ. of Ill. at Chi., *Why is Asthma Worse in Black Patients?*, SCIENCEDAILY (Jan. 6, 2017), www.sciencedaily.com/releases/2017/01/170106133056.htm [https://perma.cc/KBZ7-ZVTL].

[95] *Id.* (emphasis added).

[96] *See* Nyenhuis et al., *supra* note 79, at 4–5.

[97] *Supra* text accompanying notes 85–93.

[98] *See* Paul R. Rosenbaum & Donald B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, 70 BIOMETRIKA 41 (1983).

[99] *See* Schisterman et al., *supra* note 57, at 493 ("For estimation of total causal effects, it is not only unnecessary but likely harmful to adjust for a variable on a causal path from exposure to disease, or for a descending proxy of a variable on a causal path from exposure to disease.").

interest in this instance is the race of the subject, and the authors have adjusted for variables that measure severity of disease.[100] That severity of disease is affected by patient race is the entire premise of the paper, making this adjustment completely illogical.[101] The crude data show that blacks are observed to have more severe asthmatic disease but no difference in pattern of airway inflammation compared to whites.[102] Trying to interpret the adjusted model, we can only conclude that if blacks could be forced to have equal severity of disease, then their pattern of airway inflammation would have to become more eosinophilic. This in no way corresponds to the scientific question that was posed, nor to any other question of potential interest.[103]

Consider this analogy. Suppose that we are interested in why whites have a higher acceptance rate for admission to Harvard University than blacks, and we want to know if this is because of a racial disparity in academic performance. In our crude data we happen to observe no correlation between race and academic performance, which would seem to provide a negative answer to our original question. But instead of stopping there, we make a statistical model in which we adjust the relationship between race and academic performance for attendance at Harvard, where blacks are greatly underrepresented. Among the students attending Harvard, we find that it would have to be the case that whites must have had higher academic performance, since this is the only way for the statistical model to boost their predicted admission status above that of blacks. Now imagine that we publish a paper claiming to have discovered that whites have higher academic performance and this explains their greater numbers in the Harvard class, even though this was not true in the data. This would be incredibly disingenuous, and yet it is exactly what these authors have done in the asthma paper.[104] Sadly, this is not an atypical example.

---

[100] *See supra* text accompanying notes 89–90.

[101] Nyenhuis et al., *supra* note 79, at 2; *see also* Jay S. Kaufman & Richard S. Cooper, *Commentary: Considerations for Use of Racial/Ethnic Classification in Etiologic Research*, 154 AM. J. EPIDEMIOLOGY 291, 293 (2001); *see also supra* text accompanying notes 91–94.

[102] *See* Nyenhuis et al., *supra* note 79, at 4, Table 1; *see also id.* at 7 ("[O]ur analysis relied mostly on a single assessment of airway inflammation using induced sputum. Although induced sputum is a direct and noninvasive measure of airway inflammation, other measures, including blood eosinophil counts, exhaled nitric oxide levels, and total serum IgE levels, have been associated with greater asthma severity, yet have not consistently been shown to predict ICS treatment responsiveness to the same degree as sputum eosinophils have. Blood eosinophil counts and total serum IgE levels were measured in a subset of patients in this analysis. No difference was found in blood eosinophil counts, although African American subjects had significantly higher total serum IgE levels compared with white subjects. Larger studies examining additional measures of airway inflammation, such as blood eosinophil counts, activated airway eosinophil counts, exhaled nitric oxide levels, and total serum IgE levels, should be pursued to fully address airway inflammatory differences that might exist in African American and white patients with asthma.").

[103] *See* Nyenhuis et al., *supra* note 79, at 2.

[104] A numerical example of this phenomenon may make the point more concretely. For the 3 variables race (1=black; 0=white), high vs low academic performance (1 vs. 0) and admission to Harvard (1 vs. 0) in that order, suppose that the 8 cells of the 2x2x2 tables are: 1,1,1=20; 1,0,1=5; 1,1,0=20; 1,0,0=20; 0,1,1=44; 0,0,1=20; 0,1,0=20 and 0,0,0=20. In this case, high academic performance is a cause of getting into Harvard (OR=2.6), Black race impedes Harvard admission (OR=0.4), and academic performance is completely uncorrelated with race (OR=1). Nonetheless, when one considers only those students who are attending Harvard, then odds of high academic performance must by eighty percent higher in whites to explain their excess. In the causal inference literature, this phenomenon is known as "collider stratification bias." *See* Stephen R. Cole et al., *Illustrating Bias Due to Conditioning on a Collider*, 39 INT'L J. EPIDEMIOLOGY 417 (2010).

V. DISCUSSION

What have we learned from thinking carefully through the nature of statistical adjustment and reviewing these two examples? Even before considering adjustments, crude contrasts require various analytic decisions, such as whether rates are to be contrasted as differences or as ratios. These can produce profoundly different patterns, leading some critics to complain that the nature of a disparity is inherently ambiguous even in the real world.[105] Once we move from the real world to the hypothetical world, however, the number of decisions becomes overwhelming, and yet the basis on which these are made is seldom articulated clearly in published reports. Whether the particular imaginary population invoked by the adjustments is really optimal for some particular scientific or policy question is too rarely challenged. Most published studies, like the asthma study discussed above, list a set of covariates for adjustment without any explanation offered for why these are included and why others are excluded.[106] In the statistical literature, there do exist formal guidelines for obtaining an optimal covariate set, although there is not universal agreement on how these criteria are best operationalized.[107] The derivation of these principles often rests on the analogy of an experiment, in which individuals are randomly assigned to receive one treatment or another, and therefore presents some difficulties for non-manipulable characteristics such as race.[108] Various authors have discussed the interpretation of adjusted racial disparities models, but no broad consensus has emerged in practice.[109]

For the consumers of published findings, whether journalists or health professionals or policy makers, the adjustment process is too often seen as a mysterious "black box" that is seldom pried open. The adjusted results are considered somehow stronger, or more credible than the crude results, perhaps based on broad appreciation for the mantra that "association is not causation".[110] Sadly, there seems to be a reflexive acceptance of the notion that adjusted association *is* causation, or at least some kind of improvement. There is not nearly enough appreciation for the myriad ways in which adjustment can lead one further astray.[111] More importantly, there is not nearly enough attention to using adjustment purposefully to create the imaginary world that is most relevant to the scientific question, and thus to a recognition that these adjustments are those that would pertain in a world that is not our real world. Too often, as in the two examples shown above, the language used implies that the adjusted estimate is what was observed, when in fact the truth is that the adjusted estimate is what *would have been observed* under some imaginary circumstances.[112] It is the responsibility of the author to justify these imaginary circumstances as more scientifically relevant than those of the real world.

---

[105] *See, e.g.*, James P. Scanlan, *Can We Actually Measure Health Disparities?,* 19 CHANCE 47 (Mar. 1, 2006).

[106] *See* Nyenhuis et al., *supra* note 79, at 9.e4; *see also* Jay S. Kaufman & Richard S. Cooper, *Commentary: Considerations for Use of Racial/Ethnic Classification in Etiologic Research*, 154 AM. J. EPIDEMIOLOGY 291, 293–94 (2001).

[107] *See* Tyler J. VanderWeele & Ilya Shpitser, *On the Definition of a Confounder*, 41 ANNALS STAT. 196 (2013).

[108] *See* Jay S. Kaufman & Richard S. Cooper, *Seeking Causal Explanations in Social Epidemiology*, 150 AM. J. EPIDEMIOLOGY 113 (1999).

[109] *See* Jay S. Kaufman, *Race: Ritual, Regression, and Reality*, 25 EPIDEMIOLOGY 485 (2014); Tyler J. VanderWeele & Whitney R. Robinson, *On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables*, 25 EPIDEMIOLOGY 473 (2014).

[110] VanderWeele & Robinson, *supra* note 109, at 474.

[111] *See* Schisterman et al., *supra* note 57.

[112] *See, e.g.,* Beavis et al., *supra* note 63; *see also, e.g.,* Nyenhuis et al., *supra* note 79.

The danger of reflexively prioritizing adjusted estimates is that we may become so invested in their imaginary world that we risk losing any anchor in reality, forgetting that ultimately it is the real world that we care about. For example, Forouzanfar and colleagues recently published modeled estimates of blood pressure, hypertension burden, and their contributions to mortality for every country in the world in 1990 and 2015.[113] This took an enormous modeling effort because for the vast majority of the world's populations, there is no basic population surveillance of blood pressure or cardiovascular events (the technical details of this estimation occupy over 100 appendix pages).[114] So powerful was this model that the authors even specified the mean blood pressure in 1990 for countries that did not even exist at that time, such as Montenegro and South Sudan.[115] That a statistical model can tell us the average blood pressure in a population that has not yet come into existence is an impressive feat, but one that highlights the abstract nature of such an exercise. For example, there can be no empirical verification or refutation of the estimated value for an imaginary population.[116]

An even more impressive example is provided by Corona et al, who calculated the genetic risk for 102 diseases in 51 worldwide populations, based on 1,032 genotyped individuals.[117] The authors reported that risk of diabetes, for example, was highest in sub-Saharan Africa and lowest in Pacific Islanders.[118] As noted by the authors themselves, however, these estimates are exactly opposite of what is observed in the real world, where sub-Saharan Africans have the lowest risk of diabetes and Pacific Islanders have the highest.[119] Criticized for this complete disconnect between the modeled results and the reality of disease distribution, the authors were adamant in defense of their findings. "[W]e should not expect genetic risk to match observed risk," they wrote indignantly.[120] "The primary objection you raise to this research is based on a common misconception that observed risk and genetic risk of complex disease are expected to agree . . . . [Our research requires] neither empirical evidence about disease occurrence nor other methodology grounded in epidemiology. Inclusion of such data would be gratuitous and unnecessary. . . ."[121] These authors' devotion to the alternative reality of their model is so fierce that they claim to require no empirical evidence from the real world at all.[122] Of course they use DNA from the real world in

---

[113] Mohammad H. Forouzanfar et al, *Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015*, 317 JAMA 165 (2017).

[114] *Id.*

[115] *Id.* at app. 36, 90–91.

[116] For examples of statistical estimation for an entirely counterfactual entity, see Alberto Abadie et al., *Comparative Politics and the Synthetic Control Method*, 59 AM. J. POL. SCI. 495 (2015). This estimate can never be refuted with observed data because it pertains to an entity that is counter to historical fact.

This also invokes the philosophical debate about the demarcation between science and metaphysics. Positivist philosophers proposed that what distinguishes "science" is that its explanatory theories must be refutable based on observable evidence, but imaginary countries are never subject to direct observation. Sven Ove Hansson, *Science and Pseudo-Science*, Stanford Encyclopedia of Philosophy (Edward N. Zalta ed., 2017), https://plato.stanford.edu/entries/pseudo-science/.

[117] Erik Corona et al., *Analysis of the Genetic Basis of Disease in the Context of Worldwide Human Relationships and Migration*, PLOS GENETICS (May 23, 2013), http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003447 [https://perma.cc/TE53-CNYA].

[118] *Id.*

[119] *Id.*

[120] Ecoronap, Comment to *Analysis of the Genetic Basis of Disease in the Context of Worldwide Human Relationships and Migration*, PLOS GENETICS (July 6, 2013, 1:22 PM), http://journals.plos.org/plosgenetics/article/comment?id=10.1371/annotation/5dcab322-d620-4f9e-bfb0-6383bd42be9d [https://perma.cc/TE53-CNYA].

[121] *Id.*

[122] *Id.*

their model, but they are using this to predict disease status, and they assert that knowing who actually gets disease in the real world would be "gratuitous and unnecessary."[123] This is a remarkable perspective on science, and ought to lead to sober contemplation of what we are really hoping to accomplish with our research products.

The revered statistician George E.P. Box is well known for having observed that all models are wrong, but that some are useful.[124] If we are to be scientists who aspire to have some positive impact on the real world in which we live, it is incumbent on us to articulate exactly how our models are useful. Since models inevitably describe imaginary circumstances, it is our responsibility to draw the connection back to reality, i.e. motivating and justifying adjustments for what they tell us about the real world, not what they tell us about the world of the model in which none of us live. Making the explicit justification for why a model is useful will often highlight an additional insight, which is *for whom* the model is useful. The infinitude of possible worlds that models can occupy is limited only by our imaginations and credulities. How we arrive at one specific choice from among this limitless array is often a question of ethics and power, just as it is in so much of statistics and science.[125]

[123] *Id.*

[124] G.E.P. Box, *Robustness in the Strategy of Scientific Model Building*, *in* ROBUSTNESS IN STATISTICS 201, 202–03 (Robert L. Launer & Graham N. Wilkinson eds., 1979).

[125] *See* Steve Wing, *Whose Epidemiology, Whose Health?*, 28 INT'L J. HEALTH SERVS. 241 (1998).

APPENDIX[126]

Table 1: Population and Deaths by Age in 1970 for White Females in Miami, Alaska and the US

	MIAMI			ALASKA			UNITED STATES		
Age	Pop	Deaths	Rate	Pop	Deaths	Rate	Pop*	Deaths*	Rate
<15	114350	136	1.19	37164	59	1.59	23961	32	1.34
15-24	80259	57	0.71	20036	18	0.90	15420	9	0.58
25-44	133440	208	1.56	32693	37	1.13	21353	30	1.40
45-64	142670	1016	7.12	14947	90	6.02	19609	140	7.14
65+	92168	3605	39.11	2077	81	39.00	10685	529	49.51
	562887	5022	8.92	106917	285	2.67	91028	740	8.13

* In thousands

Rates are deaths per 1000

[126] VICTOR J. SCHOENBACH & WAYNE D. ROSAMUND, UNDERSTANDING THE FUNDAMENTALS OF EPIDEMIOLOGY: AN EVOLVING TEXT 132 (2000) (ebook) (modified from original table).