Incentivizing Access Barriers: The Unintended Consequences of Using Predicted Costs to Measure Medical Needs

RIYAD A. OMAR*

* J.D., Stanford Law School; B.A., University of Southern California. Mr. Omar's work seeks to advance best practices in healthcare innovation through accountability-by-design approaches. Previously, Mr. Omar was Vice President of Privacy and Security for Clarify Health Solutions, a provider of healthcare analytics, and Senior Vice President, General Counsel, Secretary, and Chief Strategy Officer of Practice Fusion, a leading electronic health record technology developer. Mr. Omar thanks Rebecca Payton, Courtney Morgan Crocker, Dairanetta Spain, and John Taylor, of *The University of Memphis Law Review*, for editing the piece and providing invaluable feedback to the author.

I. INTRODUCTION

Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations ("*Dissecting Bias*")[1] is part of a growing body of research warning about the propensity of big-data algorithms to perpetuate racial biases.[2] *Dissecting Bias* studied how "one of the largest and most typical examples of a class of commercial risk-prediction tools"[3] assigned risk scores to Black and White patients. The study compared those scores to a measure of the same patients' chronic conditions. Based on this comparison, the study concluded that the "widely used algorithm . . . exhibits significant racial bias."[4]

Algorithms used in decision-making have received increasing attention. Such algorithms are developed for, and used in, making a wide range of decisions once made by humans, including in policing, criminal justice, human resources, assessing creditworthiness, and a large number of other tasks.[5] This is done despite the fact that the entities who delegate their decision-making to those algorithms often do

---

1. Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 447 (2019) [hereinafter *Dissecting Bias*], https://science.sciencemag.org/content/366/6464/447.full.pdf.

2. *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 673 (2016); Andrew G. Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109, 1148 (2017); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1025 (2017) (reviewing FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015)); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 684–88 (2017); Kimberly A. Houser, *Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making*, 22 STAN. TECH. L. REV. 290, 294 (2019); Rashida Richardson et al., *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. 192, 205–07 (2019).

3. *See Dissecting Bias*, *supra* note 1, at 477.

4. *Id.*

5. *Id.* at 7 ("[C]redit-scoring algorithms predict outcomes related to income, thus incorporating disparities in employment and salary. Policing algorithms predict measured crime, which also reflects increased scrutiny of some groups. Hiring algorithms predict employment decisions or supervisory ratings, which are affected by race and gender biases. Even retail algorithms, which set pricing for goods at the national level, penalize poorer households, which are subjected to increased prices as a result.").

not fully understand how they work.[6] Further, there are times when developers may not fully understand how their algorithms operate.[7] Complicating the matter is evidence that such algorithms have a propensity to reflect societal biases, including biases based on race, ethnicity, and gender.[8] When such algorithms influence decisions made about individuals, they become instruments to perpetuate those biases.[9]

Healthcare has embraced this broad trend towards automation. What sets healthcare apart from other industries, however, are the many safeguards healthcare organizations routinely deploy to detect and remedy operational defects and wrongdoing. Large healthcare organizations, for example, often deploy quality management systems where operational requirements are documented and undergo quality assurance.[10] Many are also required to have mature data protection and

---

6. *See* discussion *infra* Section IV.C.

7. Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 638 (2017) ("Machine learning . . . is particularly ill-suited to source code analysis because it involves situations where the decisional rule itself emerges automatically from the specific data under analysis, sometimes *in ways that no human can explain.*" (emphasis added)).

8. *See, e.g.*, Barocas & Selbst, *supra* note 2, at 674–675 ("Algorithms could exhibit these tendencies even if they have not been manually programmed to do so, whether on purpose or by accident. Discrimination may be an artifact of the data mining process itself, rather than a result of programmers assigning certain factors inappropriate weight . . . Each of these steps creates possibilities for a final result that has a disproportionately adverse impact on protected classes, whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors. Even in situations where data miners are extremely careful, they can still effect [sic] discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes.").

9. *See, e.g.*, Richardson et al., *supra* note 2, at 222 ("[W]hen the dirty data generated by dubious calls for services and subjective police enforcement and reporting is used in predictive policing systems, the technology can produce predictions that further perpetuate confirmation feedback loops.").

10. *See*, *e.g.*, Ian R. Lazarus & M. Weston Chapman, *"ISO-Style" Healthcare: Designed to Keep Patients, Practitioners and Management Safe*, BECKER'S HOSP. REV. (Sept. 26, 2013), https://www.beckershospitalreview.com/hospital-management-administration/iso-style-healthcare-designed-to-keep-patients-practitioners-and-management-safe.html (noting that over 5,000 hospitals and 10,000 other institutions are accredited as implementing a quality management system); 45 C.F.R. § 170.315(g)(4) (2015) (requiring that all capabilities of electronic health record technology that are certified for use in federal healthcare quality improvement programs

compliance operations[11] that are well-positioned to review algorithms to ensure that their specifications comply with applicable laws.[12] Hospitals and health systems frequently have medical officers who oversee clinical operations and are versed in a wide range of issues surrounding healthcare delivery, including the fact that minority groups face significantly greater barriers to accessing medical care.[13] They also have research and data science expertise that can be utilized to validate how algorithms function before and after deployment in clinical operations. It may seem a remote probability, therefore, that an algorithm that "exhibits significant racial bias"[14] would be deployed "industry-wide . . . affecting millions of patients . . . ."[15]

This may be why *Dissecting Bias*'s findings received prompt regulatory attention. On the day *Dissecting Bias* was published, New York's Departments of Financial Services and Health wrote to UnitedHealth Group ("New York Letter*"*), corporate parent of Optum, the developer of Impact Pro, one of the leading health risk scoring algorithms.[16] Although *Dissecting Bias* does not identify Optum or Impact Pro by name, New York cites the study as "show[ing] that Impact Pro's flawed algorithm ranked healthier [W]hite patients as equally at risk for future health problems . . . as [B]lack patients who suffered from

---

were developed, tested, implemented and maintained in conformance with a quality management system ("QMS") established by the Federal government or mapped to such QMS).

11. *See, e.g.*, 45 C.F.R § 164.530(c)(1) (2015) ("A covered entity [health care provider, health plan, or health care clearinghouse] must have in place appropriate administrative, technical, and physical safeguards to protect the privacy of protected health information.").

12. *See, e.g.*, 45 C.F.R. § 164.308(a)(4)(i) (2015) (stating that healthcare organizations must "[i]mplement policies and procedures for authorizing access to electronic protected health information that are consistent with the applicable requirements of [HIPAA's Privacy Rule]").

13. *See, e.g.*, *100 Hospital and Health System CMOs to Know*, BECKER'S HOSP. REV. (Feb. 19, 2020), https://www.beckershospitalreview.com/lists/100-hospital-and-health-system-cmos-to-know-2020.html.

14. *Dissecting Bias*, *supra* note 1.

15. *Id.*

16. Letter from the New York State Dep't of Fin. Servs. and the New York State Dep't of Health to David S. Wichmann, CEO, UnitedHealth Group Inc. 1 (Oct. 25, 2019) [hereinafter New York Letter], https://dfs.ny.gov/system/files/documents/2019/10/20191025160637.pdf.

far more chronic illnesses."[17] New York calls on UnitedHealth to "cease using Impact Pro (or any other data analytics program) if [it] cannot demonstrate that it does not rely on racial biases or perpetuate racially disparate impacts."[18]

The New York letter raises a second concern. "In addition," it notes, "we are troubled by the potential for conflicts of interest to the extent that *the entity that controls the algorithm also is affiliated with both providers and insurers*, and the *algorithm considers costs*."[19] The observed racial bias could be a side effect of prioritizing the *financial interests* of UnitedHealth over the *health needs* of patients. Patients who currently have trouble accessing medical care, for example, will generate *increased* medical costs if the algorithm recommends that those patients receive enhanced medical attention. Insurers, therefore, financially benefit if the algorithm disregards those patients because their pre-existing barriers to care already effectively reduce their costs.

The New York letter's assignment of accountability is refreshingly straightforward. It sidesteps quandaries regarding the mysteries of algorithm development by imposing familiar principles of accountability for what algorithms do and how they are used. New York's laws govern the activities of entities "irrespective of whether they themselves are collecting data . . . or using and developing algorithms or predictive models . . . ."[20] This assignment responsibility, however, has implications for how we interpret *Dissecting Bias*'s findings.

Based on the facts laid out in the study, for example, there is a strong possibility that the "significant racial bias" arose from *how* an otherwise *unbiased* algorithm was implemented. Further, it is also possible that the specified implementation included countermeasures designed to mitigate potential "racially disparate impacts" that could arise from the algorithm's natural operation.

Neither possibility, however, lessens the concerns raised by *Dissecting Bias*. If the significant bias arose from the implementation, for example, it is important to know what deliberations, if any, occurred. If safeguards were deployed to counter any potential "racially disparate impacts," it is important to identify those safeguards and understand

17. *Id.*
18. *Id.*
19. *Id.* (emphasis added).
20. *Id.*

how successful they were. Finally, it is important to understand whether those safeguards are universally applied whenever the algorithm is used to influence medical decision-making. If *Dissecting Bias* is correct that these implementations are "typical of [an] industry-wide approach"[21] and "applied to roughly 200 million people in the United States"[22] it is important to have confidence that they do not give rise to the concerns raised by New York's regulators.

This Article evaluates *Dissecting Bias*, raising questions about whether the observed bias arose from the algorithm itself, its implementation, or may be an artifact of study design. Each possibility, however, reflects risk management gap in how decision-support algorithms are currently implemented in healthcare. This gap can readily result in an otherwise "unbiased algorithm" being used to "perpetuate racial biases." It could just as easily convert an accurate cost-prediction algorithm into a means of denying or de-prioritizing medical care for patients already suffering medical access burdens. This Article argues that these risks could be significantly mitigated if healthcare organizations applied familiar data governance principles in how they manage the algorithms they use to support clinical decision-making.

This Article's discussion is divided into five parts, the first of which is this introduction. Part II examines risk scores and predictive models. It also discusses how their predictive accuracy is measured, as well as common risks arising from their development and implementation. Part III discusses *Dissecting Bias* and the concerns raised by the New York regulators. Here the focus is on how an inattentive implementation of a cost-prediction algorithm could adversely impact patients who are already burdened by medical access barriers, including those from racial and ethnic minority groups. Part IV examines whether the observed "significant racial bias" described in *Dissecting Bias* was the result of the algorithm's design or its implementation. Part IV identifies the many *prima facie* risks inherent in the implementation described in *Dissecting Bias*, as well as strategies available to address those risks. Finally, Part V discusses how familiar principles of data governance can be used to address many of the inherent risks associated with the use of algorithms in medical decision-making.

---

21. *See Dissecting Bias*, *supra* note 1.

22. *Id.*

II. THE USE AND MISUSE OF PREDICTIVE MODELS

A. Risk Scores in Health Care

A "risk score" is a tool for estimating the severity or likelihood of an adverse event occurring.[23] The Framingham Risk Score ("FRS"), for example, estimates the risk of a patient developing cardiovascular disease ("CVD") within ten years based on a patient's age, gender, cholesterol, systolic blood pressure, HDL-C count, and whether the patient is a smoker or diabetic.[24]

Risk scores support decision-making in at least two important respects. First, by design, they are intended to be easy to use.[25] In FRS, for example, a doctor can add up the points attributable to each of the model's *inputs*—a patient's age, gender, smoking status, etc.—in order to obtain an overall score—the *output*—estimating the risk that the patient will develop CVD within ten years.[26] In FRS, each of the risk factors contributes points to an aggregate score that can range from -3 or less, on the low end, to 21 or more on the high end.[27] Scores for male and female patients are then partitioned into their respective "10-Year CVD Risk," which for women is "high" at 21.5 or higher, and "medium" between 10 and 18.5.

The second way FRS supports decision-making is that it provides clarity between its medical *inputs*—for example, age, gender, smoking, diabetes—and its *outputs*, its risk score. This clarity allows doctors to understand the impact that each factor contributes to the ultimate risk score. In male patients, for example, being a smoker contributes four points to his overall FRS. This makes FRS "easy to use"

---

23. *See* Berk Ustun & Cynthia Rudin, *Learning Optimized Risk Scores*, 20 J. MACHINE LEARNING RSVH. 1,1 (2019), https://jmlr.org/papers/volume20/18-615/18-615.pdf.

24. *See, e.g.*, *Framingham Risk Score (FRS) Estimation of 10-year Cardiovascular Disease (CVD) Risk*, CANADIAN CARDIOVASCULAR SOC'Y (2017) [hereinafter *Framingham Risk Score*], https://ccs.ca/app/uploads/2020/12/FRS_eng_2017_fnl_greyscale.pdf.

25. Ustun & Rudin, *supra* note 23, at 59 ("[R]isk scores are used because they are easy to use, understand, and validate.").

26. *See, e.g.*, *Framingham Risk Score*, *supra* note 24.

27. *Id.*

for its intended purpose, which is to aid physicians in assessing and explaining medical risks for and to their patients.

The utility of risk scoring has extended to other endeavors, notably in financial services, such as credit scoring, and in the criminal justice system—scoring the risk of recidivism in the context of bail determinations, sentencing, and parole.[28] In healthcare, risk scoring has expanded as well. Risk scores are now used for "risk adjustment," where transfer payments among insurers or between insurers or providers are based, in part, on the measure of financial risk that a group of patients or a type of bundle of services presents via a "risk score."[29] Similar financial risk scoring approaches are utilized in the financial administration of healthcare, including in financial modeling and resource allocation, program evaluation, provider or health plan reimbursement, pricing health plans, and projecting future claims costs.[30]

Because the purpose of a financial risk score is very different from that of a medical event risk score (such as FRS), financial risk scores are often developed and used very differently. The 2012 risk adjustment model of the Centers for Medicare & Medicaid Services ("CMS"), for example, calculates an enrollee's risk score based on her demographic and health status information *relative to the average expenditure* for each enrollee.[31] If, for example, a 57-year-old woman with a specified medical condition is expected to incur $1,200 in medical costs, and the average medical cost for an enrollee is $1,000, then

---

28. *See* Ustun & Rudin, *supra* note 23, at 1–2.

29. *See, e.g.*, GEOF HILEMAN ET AL., SOCIETY OF ACTUARIES, RISK SCORING IN HEALTH INSURANCE: A PRIMER 4 (2016) ("The goal of the risk adjustment program is to adjust payments to insurers to reflect the actual risk profile of the individuals who enroll in their plans relative to other plans in the same state and block. The risk adjustment program is divided into two stages. The first stage is the determination of a 'risk score' of each insured population. The second stage is the risk transfer formula that is used to balance the premiums among the health plans to reflect differences in risk scores of the enrolled population by health plan.").

30. IAN DUNCAN, HEALTHCARE RISK ADJUSTMENT AND PREDICTIVE MODELING 14 (2d ed. 2018) (providing overview of the use of prediction models in healthcare financial applications).

31. CTRS. FOR MEDICARE AND MEDICAID SERVS., DEP'T OF HEALTH & HUM. SERVS., HHS RISK ADJUSTMENT MODEL 7 (2012), https://www.cms.gov/CCIIO/Resources/Files/Downloads/fm-1c-risk-adj-model.pdf (describing how "[e]ach enrollee risk score is based on the individual's demographic and health status information . . . [c]alculated relative to average expenditures").

the patient's risk score would be 1.2.[32] The "severity" of the patient's risk score, therefore, is not solely based on the severity of her medical conditions. Rather, it is based on its *comparative severity* in relation to other patients. This contrasts with FRS, in which the severity of the patient's FRS is based *solely* on her own health attributes—her age, cholesterol, systolic blood pressure, HDL-C count, diabetes diagnosis, and smoking status.[33]

Second, the key performance indicator of a financial risk score is its ability to make accurate financial predictions. The patient's medical condition is relevant only for its contribution to accurately anticipating the medical costs associated with that patient. Medical risk scores, by contrast, are primarily interested in the medical condition itself. FRS, for example, does not consider medical costs in predicting the likelihood that a patient will be diagnosed CVD within 10 years. FRG is solely interested in the factors that contribute to CVD.

Third, although financial risk scores can be—and often are—applied to individuals, their ultimate effectiveness is often measured in how effectively they apply to populations or pools of risk:

> The risk of a population will be different than that of an individual, because of the "spread of risks" that is inherent when a number of lives are pooled in a population. An individual may be highly risky because of his condition-based risk or lifestyle risk factors, yet the population of which he is part may not represent a significant risk.[34]

Thus, although a cost prediction may be applied to each individual in a given risk pool, it is not a flaw in the algorithm if its predictions are wrong about many (or even most) of the patients in that pool. What is often most important is how accurate the model is when all of the individual predictions are combined into a composite prediction. The accuracy of that composite prediction is what, for example, insurers are

---

32. *Id.*

33. In terms of "ease of use," it is worth noting that the form factor of this risk score is related to its intended purpose. Here, the score "1.2" is easy to plug into a formula used to calculate a financial payment. This contrasts with FRS, where the intended use is to aid clinicians in performing a clinical assessment based on inputs they can directly observe in a manner that reflects the comparative contribution of each input.

34. DUNCAN, *supra* note 30, at 7.

often most interested in. This contrasts with medical risk scores, like FRS, where a patient is primarily interested in its accuracy with respect to her alone.

B. Measuring Predictive Accuracy

1. Predictions About Groups

These factors have a significant impact on how the accuracy of financial risk scores are measured. "One of the more useful measures of predictive fit" is to conduct a statistical analysis that compares an algorithm's cost *predictions* for categories of patients sharing common characteristics to their *actual* costs.[35] Here "predictive fit" is measured by how closely those predictions accurately forecasted the costs.

In its 2016 report, Accuracy of Claims-Based Risk Scoring Models ("SOA Report"), [36] the Society of Actuaries ("SOA") represented this difference as a "predictive ratio" in which an algorithm's accuracy is expressed as a percentage equal to the predicted costs divided by the actual costs. Accordingly, "[a] predictive ratio in excess of 100 percent indicates that a model *overestimates* the risk level for that group, while a predictive ratio below 100 percent indicates that the model *underestimates* the risk level."[37] An example can be seen in Table 4.4.2 of the SOA Report, which shows the predictive ratios for one million data subjects[38] for a number of actuarial risk scoring algorithms,[39] including:

35. *See* GEOF HILEMAN & SPENSER STEELE, ACCURACY OF CLAIMS-BASED RISK SCORING MODELS 24 (2016), https://www.soa.org/globalassets/assets/files/research/research-2016-accuracy-claims-based-risk-scoring-models.pdf ("One of the more useful measures of predictive fit is the predictive ratio, defined here as the mean risk score for a group of individuals divided by the mean actual scaled cost for that same group.").

36. *Id.*

37. *Id.* (emphasis added).

38. *Id.* at 11–12, 26 (discussing the characteristics of the study sample).

39. *Id.* at 73.

Table 1

Predictive Ratios by Health Conditions (Prospective; Offered Weights; No Censoring)						
	Heart Disease	Mental Illness	Diabetes	Low Back Pain	Asthma	Arthritis
Diagnosis-and-Pharmacy Models						
Impact Pro	59.9%	78.3%	85.0%	76.4%	78.2%	74.7%

As noted in the SOA Report, a "predictive ratio within plus or minus 10 percent of 100 percent indicates a reasonable degree of accuracy for a subgroup."[40] *None* of the algorithms evaluated in the SOA Report hit this performance target in predicting the costs of condition-based sub-populations. Across the industry, algorithms systematically underestimated the medical costs associated with sub-populations defined by their specified medical conditions.

Conversely, all vendors fared significantly better at predicting costs based on a demographic group's gender and age. Impact Pro, for example, hit the plus or minus 10% performance target with *every* subgroup in most of its configurations, as is shown below.[41]

Table 2

Predictive Ratios by Age-Sex, Prospective Models (Offered Weights; $250,000 Censoring)						
	Children 0-6	Children 7-18	Males 19-44	Males 45-64	Females 19-44	Females 45-64
Diagnosis-Only Models						
Impact Pro	116.0%	107.9%	102.1%	99.5%	101.6%	95.4%
Pharmacy-Only Models						
Impact Pro	111.2%	92.9%	99.0%	100.9%	100.0%	100.5%
Diagnosis-and-Pharmacy Models						
Impact Pro	100.7%	99.9%	101.8%	100.1%	102.4%	97.8%

On the other extreme, all algorithms fared poorly at predicting both "low cost" and "high cost" demographic sub-groups, significantly

40. *Id.* at 24–25 (citation omitted).

41. *Id.* at 30.

underestimating the costs of "high cost patients," and significantly overestimating the cost of "low cost patients."[42] All of them, for example, had results that looked similar to Impact Pro's depicted below.[43]

Table 3

Predictive Ratios by Cost Percentile (Prospective; Offered Weights; No Censoring)								
	0-20th	20-40th	40-60th	60-80th	80-90th	90-95th	95-98th	98-99th
Diagnosis-and-Pharmacy Models								
Impact Pro	8420%	686%	340%	189%	114%	74%	52%	29%

The SOA Report highlights the extent to which an algorithm's predictive accuracy depends on the question it is being asked, even with respect to the same overall sample of patients.[44] As a general rule, *all* of the vendors had similar results to those described above, and their similar performance largely reflects the reality that an algorithm's accuracy at predicting one set of variables is not indicative of its ability to predict others.[45]

2. Predictions About Individuals

Actuarial predictions about groups of people are very different than clinical instruments designed to diagnose or make predictions about an individual. Even when, for example, actuarial forecasts are applied to individuals for the purposes of calculating the financial risk of an overall group of patients, the quality of those predictions is often measured in terms of how effectively they predict the overall financial costs associated with that group. Risk scores can, in principle, be repurposed to make predictions about specific individuals, such as predicting whether a patient will be in the top 1% of most expensive patients,[46] or, in the criminal justice context, predicting whether a parolee

42. *Id.* at 77 (showing the predictive ratios by cost percentile of all models).

43. *Id.*

44. *See supra* notes 36–41 and accompanying text (demonstrating the varying predictive accuracy of algorithms).

45. *See generally* HILEMAN & STEELE, *supra* note 35, at 16–46 and 59–80 (showing the results of the study).

46. *Id.* at 43 ("In addition to the calculation of the average risk score for a group of individuals, risk scoring models are also often used to identify the very highest cost individuals.").

will commit crimes in the future.[47] This, in turn, impacts how their predictive accuracy is evaluated.

A common approach is to measure the instrument's sensitivity and specificity. An instrument's sensitivity is also known as its "true positive rate" because it measures the instrument's ability to correctly identify whether an individual belongs in that group.[48] It is measured by "the percentage of individuals that are correctly identified as being among [the target group]."[49] An instrument's specificity (or "true negative rate"), on the other hand, measures its ability to correctly identify when an individual does *not* belong in the group.[50] It is measured by "the percentage of individuals that are correctly identified as *not* being in [that target group]."[51]

Frequently, an instrument can set different thresholds for when it returns a "positive" or "negative" result. A high threshold, for example, may be set when one wants to be certain that a "positive" result includes as few "false positives" as possible, such with a diagnostic test in order to avoid subjecting a patient to invasive, risky or expensive medical procedures.[52] A low threshold, on the other hand, is often appropriate as a screening mechanism, where the goal is to detect the incidence of a potential condition.[53]

The "beyond a reasonable doubt" standard in criminal justice is an example of such a high threshold. The tradeoff with that threshold is a larger number of "false negatives" that result in guilty individuals

---

47. *See* discussion *infra* Section II.C regarding the COMPAS risk scoring system.

48. HILEMAN & STEELE, *supra* note 35, at 44.

49. *Id.*

50. *See id.*

51. *Id.* (emphasis added).

52. *See* L. Daniel Maxim et al., *Screening Tests: A Review with Examples*, 26 INHALATION TOXICOLOGY 811, 813, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4389712/pdf/uiht26_811.pdf ("[H]igh specificity tests perform well for diagnosis because of low false positive errors. Tests with low specificity have the disadvantage that (among other things) many subjects without the disease will screen positive and potentially receive unnecessary (and possibly invasive, risky or expensive) follow-up diagnostic or therapeutic procedures.")

53. *Id.* ("A highly sensitive test means that there are few *false negative* results; few actual cases are missed. *Ceteris paribus*, tests with high sensitivity have potential value for screening, because they rarely miss subjects with the disease.") (second emphasis added)).

going free. Conversely, to ensure no guilty person goes unpunished, we would set a very low threshold that reduces the number of "false negatives" as much as possible. "Strict liability" is a converse example of that standard, where an organization is deemed at fault. The tradeoff in strict liability is a significant increase in "false positives" in which organizations will be held liable even in situations where they took every conceivable action to avoid a harm.

In the same way, an instrument's threshold can often be toggled to reflect a desired level of certainty about an individual's exclusion from a class, which results in a tradeoff between the desire for "true negatives" versus "false negatives."

This tradeoff between an instrument's comparative sensitivity and specificity at various confidence goals can be measured by plotting the changes in both measures as the instrument's thresholds vary.[54] The collection of these data points is called a "receiver operating characteristic" or "ROC" curve.[55] The overall accuracy of the instrument is often measured by what is called the "area under the ROC curve" or "AUC,"[56] in which an instrument's predictive ability is measured by the geometric area under the ROC curve. The total available "space" is assumed to be 1.0.[57] The AUC of any instrument will always be expressed as a decimal, such as 0.75.[58]

If one simply flipped a coin to decide whether an individual had CVD, that "medical instrument" would have an AUC of 0.5. All "satisfactory" instruments, therefore, must have AUCs *above* 0.5. How much further above 0.5 is required depends on its intended purpose. What might be suitable for diagnosing a lethal or expensive medical condition, for example, may be too stringent for a basic screening tool. As a general rule, "an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based

54. *See generally* Jayawant N. Mandrekar, *Receiver Operating Characteristic Curve in Diagnostic Test Assessment*, 5 J. THORAC. ONC. 1315, 1315–1316 (2010), https://www.jto.org/article/S1556-0864(15)30604-3/pdf (discussing the use of ROC curves in assessing the validity of diagnostic tests).

55. *Id.*

56. *See, e.g.*, *id.*

57. *See id.* at 1316.

58. *See id.*

on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding."[59]

The SOA Report provides a limited assessment of the ability to make predictions about individuals.[60] It conducted an AUC assessment for all of the major medical cost-predictive algorithms' ability to correctly identify the "top 1%" of most costly patients.[61] All of the vendors had AUCs of 0.8 to 0.9, considered "excellent" as described above.[62] Impact Pro's AUCs ranged between 0.841 and 0.871.[63] The SOA Report took a different approach to measuring patients falling outside the "top 1%." For the overall population, it assessed each algorithm based on "what percentage of individuals are predicted accurately within an absolute error of X points."[64] If, for example, the prediction was allowed to be inaccurate by "plus or minus 20%"[65] then what percentage of patients would be correctly predicted by that tolerance level?

Not surprisingly, all of the algorithms' predictions increase in "accuracy" the larger the "tolerable error" range becomes. For example, when the tolerable error rate is set at plus or minus 25%, the "age-sex" demographic statistic correctly predicts 14.3% of the patients.[66] If, on the other hand, the tolerable error rate is set to plus or minus 250%, the same model correctly predicts 94.8% of the patients.[67] The SOA Report uses this approach to create tolerance curves for each of the algorithms, where each algorithm's predictive accuracy increases as the tolerable error range increases. SOA then uses this to create a modified AUC analysis, where algorithms' predictive accuracy is

---

59. *Id.*

60. HILEMAN & STEELE, *supra* note 35.

61. *Id.* at 46.

62. *Id.* Note that patients in the "top 1%" of costs are frequently suffering long-term, high-cost conditions that span many years. Consequently, although risk scoring algorithms treat these as "predictions," they are not predictions about an uncertain future. Rather, they are often more akin to observations of an ongoing event that is subject to little contingency.

63. *Id.*

64. *Id.* at 39 (emphasis omitted).

65. *E.g.*, if the predicted costs were $100, then that prediction would be "correct" if the actual costs were $119, but "incorrect" if the actual costs were $121.

66. HILEMAN & STEELE, *supra* note 35, at 40 tbl.4.5.1 (Cumulative Distribution of Error – Concurrent Dx-Only Models).

67. *Id.*

measured by the area of the chart that falls under the aforementioned tolerance curve.[68] When, for example, the tolerable error range is set to a maximum of 100%, all of the algorithms have AUCs of approximately 50%.[69] When the tolerable error rate is set to a maximum of 300%, the algorithms' AUCs increase to the 75%–79% range.[70]

To be clear, however, the AUC assessments described above do not purport to assess the algorithms' ability to correctly predict the medical conditions of individual patients.

C. Propensity to Misuse Risk Scores

Risk scores can be very effective when used for their validated purposes. The adoption of FRS, for example, has been credited with contributing to a 67.5% decline in deaths from heart disease in the United States between 1969 and 2013.[71] At the same time, risk scores also have well-known limitations. One such limitation is inherent in their actuarial foundations—at their heart, they describe risks associated with individuals who have X attributes, *not* the risk of a specific individual. Moreover, a risk score will be wholly "blind" to factors that were not accounted for in its development. For example, the Rapid Risk Assessment for Sexual Offense Recidivism risk score used to predict sex offender recidivism would have given notorious serial killer, Jeffery Dahmer, a score of zero[72] because the risk model is "blind" to factors specific to Jeffery Dahmer, such as his "multiple decade necrophiliac obsession."[73] Moreover, as discussed above, a model's predictive validity in one specific use may say very little about the model's

68. *See id.* at 39–41 (discussion of "tolerance curves").

69. *Id.* at 43 tbl.4.5.3 (Area Under Tolerance Curves Prospective Models (Offered Weights, Uncensored)).

70. *Id.*

71. NAT'L INSTS. OF HEALTH, OFF. OF SCI. POL'Y, THE FRAMINGHAM HEART STUDY: LAYING THE FOUNDATION FOR PREVENTATIVE HEALTH 3 [hereinafter FRAMINGHAM HEART STUDY], https://www.nih.gov/sites/default/files/about-nih/impact/framingham-heart-study.pdf ("From 1969 to 2013, U.S. deaths from heart disease fell 67.5% and deaths from stroke fell 77%.").

72. Shoba Sreenivasan et al., *Actuarial Risk Assessment Models: A Review of Critical Issues Related to Violence and Sex-Offender Recidivism Assessments*, 28 J. AM. ACAD. OF PSYCHOL. & L. 438, 440 (2000), https://pubmed.ncbi.nlm.nih.gov/11196254/.

73. *Id.*

ability to predict something that may appear to be closely related. As the SOA warns, "a risk scoring model designed for one outcome . . . may not be a suitable risk scoring model for another outcome . . . ."[74]

Despite these known limitations, there is a documented propensity for risk models to be misapplied. Following hurricanes Katrina, Rita, and Wilma in 2005, for example, observers criticized risk scoring models for failing "to predict the disastrous consequences of the levees failing."[75] According to the developers, however, their models "were not intended to cover the flooding due to the prolonged pressure on the levees rather than from overtopping storm surge."[76] Clients failed to recognize the factors the models were blind to.

Additionally, even the best models have error rates. The National Center for State Courts surveyed multiple assessments of COMPAS, a risk scoring algorithm used in the criminal justice system to predict the likelihood that an individual will commit a future crime, and reporting AUC values ranging from .51 on the low end to .73 on the high end.[77] "An AUC = .5 means that an assessment tool is no better than chance at discriminating between recidivists and non-recidivists. The closer the AUC value is to 1, the more effective the assessment tool is at discriminating between recidivists and non-recidivists."[78] An algorithm that has an AUC of less than 1, therefore, will inevitably generate false positives and false negatives.[79] The COMPAS risk scoring system, for example, assigned a risk of violence score of 1—COMPAS's lowest such score—to a defendant that had fifteen prior charges, including attempted murder, aggravated battery, and carrying

74. HILEMAN ET AL., *supra* note 29, at 5.

75. INT'L ACTUARIAL ASS'N, COMPREHENSIVE ACTUARIAL RISK EVALUATION (CARE) 13 (May 2010), https://www.actuaries.org/CTTEES_FINRISKS/Documents/CARE_EN.pdf.

76. *Id.*

77. PAMELA M. CASEY ET AL., NAT'L CTR. FOR STATE COURTS, OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS app. A-23 (2014), https://www.ncsc.org/__data/assets/pdf_file/0018/26226/bja-rna-final-report_combined-files-8-22-14.pdf.

78. *Id.* at 17.

79. *See id.* (the study mentioned *supra* note 77 uses the terms "incorrect classifications" and "false alarms" to describe false positives and negatives.).

a concealed firearm.[80] That defendant was subsequently arrested for kidnapping, sexual battery, and aggravated assault with a deadly weapon, notwithstanding his risk score of 1.[81]

The propensity for sophisticated organizations, including insurers and criminal justice bodies, to misuse prediction models has prompted alarm. International Actuarial Association, for example, discussed the importance of developers and users of risk models "identify[ing] the situations and scenarios in which the model results would be unreliable," including:

> [where] [t]he data is found to be insufficiently representative of the underlying situation . . . [t]he implicit assumptions of the model that drive the formation of the formulas that make up the model are no longer valid . . . [and] [t]he explicit assumptions are no longer valid because the environment is not sufficiently similar to the situation when the assumptions were formed.[82]

Similarly, the user manual for the COMPAS risk scoring software specifically instructs its clients that "risk assessment is about predicting group behavior (identifying groups of higher risk offenders)—it is *not about prediction at the individual level*. . . . Our risk scales are able to identify groups of high-risk offenders—*not a particular high-risk individual*."[83]

Even these warnings, however, have not precluded misuse. COMPAS's developer, for example, has testified in cases to clarify the proper use of COMPAS scores; in at least one case, this led to a defendant's sentence being reduced to what the judge would have imposed absent the misapplication of COMPAS scores.[84] Wisconsin's

---

80. Cynthia Rudin et al., *The Age of Secrecy and Unfairness in Recidivism Prediction*, HARV. DATA SCI. REV., 1, 22 (2020) (chart summarizing the criminal history of "Bart Sandell," a pseudonym of a real defendant used in the report).

81. *Id.*

82. INT'L ACTUARIAL ASS'N, *supra* note 75.

83. NORTHPOINTE, PRACTITIONER'S GUIDE TO COMPAS CORE 31 (Mar. 19, 2015) (emphasis added), http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf.

84. Julia Angwin, et al., *Machine Bias*, *There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-

Supreme Court also recognized the propensity to misapply COMPAS scores. In considering their use in criminal sentencing, the Court required that all presentencing reports containing COMPAS risk scores include notices describing the risk scores' limitations, including that COMPAS scores are only "able to identify groups of high-risk offenders, not a particular high-risk individual . . . ."[85]

D. Unique Risks of Black Box Algorithms

Developing effective risk scores is an empirical endeavor. The scientific method, based on the publication of research results, is recognized as a way to rapidly finding weaknesses in the original research, thereby improving on it. FRS, for example, has been studied extensively. Thanks to this research, the FRS factors have evolved and improved "to include other risk factors, and a suite of additional risk calculators . . . [including] for heart disease, heart failure, atrial fibrillation, claudication (exercise-induced leg cramping), stroke, diabetes, high blood pressure, and more."[86]

Some developers of risk scoring algorithms, however, treat their algorithms as a trade secret, in particular the relationship between the algorithms' inputs and outputs. These are sometimes called "black box" algorithms. Black box algorithms present unique risks to customers. First, by obscuring the relationship between inputs and outputs, it is more difficult for clients to assess when the algorithm's risk score is accurate. With FRS, for example, a doctor can often immediately detect when there is a discrepancy between a patient's FRS and the factors giving rise to the risk. If a 15-year-old female, non-smoker, and non-diabetic with normal blood pressure, had a documented FRS of 25, a physician immediately would know an error had occurred.

With a black box algorithm, on the other hand, the customer does not know "how factors are weighed or how risk scores are

---

in-criminal-sentencing ("Judge Babler reduced Zilly's sentence, from two years in prison to 18 months. 'Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months . . . .'").

85. State v. Loomis, 881 N.W.2d 749, 769 (Wis. 2016).

86. *See* FRAMINGHAM HEART STUDY, *supra* note 71 ("From 1969 to 2013, U.S. deaths from heart disease fell 67.5% and deaths from stroke fell 77%.").

determined."[87] As a result, it is often difficult, if not impossible, for a user to understand when an error has occurred. When the COMPAS system generates a risk-of-violence score of 1 to a defendant that has fifteen prior charges, including attempted murder, aggravated battery and carrying a concealed firearm,[88] for example, the customer often cannot tell whether this is a reflection of the model's error rate or the result of a typographic or other error made during the inputting process.

A second consequence of the lack of transparency is that errors can persist in black box algorithms for significantly longer than in transparent algorithms. Models improve when their errors are understood and documented. When users cannot tell when they are observing an error, the model is deprived of the input needed to improve its accuracy.

E. Propensity for Algorithmic Data to Reflect Societal Biases

Because developing risk scores is an empirical endeavor, research relying on data about humans is susceptible to the same biases as any other human endeavor. Numerous studies have shown that algorithms that rely on large volumes of data about the world have a propensity to reflect biases that may exist in society.[89] *Big Data's Disparate Impact*[90] describes the following five mechanisms by which "disproportionate[] adverse outcomes might occur . . . " whenever developers mine big data to develop algorithms or models:[91] (1) defining the "target variable" and "class label";[92] (2) training data[93]; (3) data collection;[94] (4) feature selection[95]; and (5) proxies.[96]

---

87. *Loomis*, 881 N.W.2d at 769 (noting that the "proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined").

88. Rudin et al., *supra* note 80, at 22.

89. *See, e.g.*, Barocas & Selbst, *supra* note 2, at 673 (studies listed in the article conclude that biases can be shown through algorithms).

90. *Id.*

91. *See id.* at 677–92 (discussing five-step mechanism that influences disproportionate outcomes in algorithms).

92. *See generally id.* at 677–80.

93. *Id.* at 680–84.

94. *Id.* at 684–87.

95. *Id.* at 688–90.

96. *Id.* at 691–92.

All of these mechanisms introduce risks of potentially replicating societal biases. Three of those mechanisms—defining the target variable, training data and data collection—are relevant to the discussion in Part III below.

When developing an algorithm, "[t]he outcomes of interest . . . are known as target variables."[97] "The proper specification of the target variable," however, "is frequently not obvious."[98] A health insurer, for example, may be interested in patients with multiple chronic conditions because it wants to reduce costs associated with avoidable medical events, such as hospitalizations. From the insurer's perspective, the target variable for those patients may be defined as "potentially very expensive" patients. For a physician, on the other hand, those same individuals are patients with specific medical conditions. From the physician's perspective, the target variable may be "multiple chronic conditions." Although there may be significant correlation between these two targets, they are nevertheless measured differently and reflect distinct priorities. It is in this initial phase of an algorithm's development that it can be set in a manner that inadvertently perpetuates societal biases. *Big Data's Disparate Impact* offers a good illustration of how this can occur:

> St. George's Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants based on its previous admissions decisions. Those admissions decisions, it turns out, had systematically disfavored racial minorities and women with credentials otherwise equal to other applicants. In drawing rules from biased prior decisions, St. George's Hospital unknowingly devised an automated process that possessed these very same prejudices. As editors at the British Medical Journal noted at the time, '[T]he program was not introducing new bias but merely reflecting that already in the system.'[99]

---

97. *Id.* at 678.

98. *Id.*

99. *Id.* at 682 (citing Stella Lowry & Gordon Macpherson, *A Blot on the Profession*, 296 BRIT. MED. J. 657, 657 (1988)).

Bias can also arise from anomalies in the data being mined. As noted in *Big Data*:

> [B]iased training data leads to discriminatory models. This can mean two rather different things, though: (1) if data mining treats cases in which prejudice has played some role as valid examples to learn from, that rule may simply reproduce the prejudice involved or (2) if data mining draws inferences from a biased sample of the population, any decision that rests on these inferences may systematically disadvantage those who are under– or overrepresented in the dataset.[100]

In the context of healthcare algorithms, there are numerous data quality challenges. First, "[d]ata-entry errors are a serious problem for medical records."[101] Moreover, as discussed in Part III below, certain demographic subgroups, including protected classes, may experience barriers to accessing medical care. These barriers, in turn, can result in deficits in the quantity, quality, and accuracy of medical information that has been collected about them. Regardless of how the algorithm was trained, biased data collection can impact an algorithm's operation if the data it relies on for its ongoing function reflects societal biases. As noted in *Big Data's Disparate Impact*:

> Decisions that depend on conclusions drawn from incorrect, partial, or nonrepresentative data may discriminate against protected classes . . . . [T]he quality and representativeness of records might vary in ways that correlate with class membership (e.g., institutions might maintain systematically less accurate, precise, timely, and complete records for certain classes of people). Even a dataset with individual records of consistently high quality can suffer from statistical biases that fail to represent different groups in accurate proportions.[102]

---

100. *Id.* at 680–81 (citing Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 3, 20 (Bart Custers et al. eds., 2013)).

101. Rudin et al., *supra* note 80, at 22.

102. *See* Barocas & Selbst, *supra* note 2, at 684.

These mechanisms are inherent in any algorithm that relies on large volumes of real-world information, whether that is in its initial development or in its ongoing operation. Accordingly, these mechanisms give rise to a natural propensity for such algorithms to reflect biases that are latent in the data they process. Absent deliberate steps specifically directed at identifying and remediating such biases, it is often more likely than not that biases reflected in society will be reflected in the operation of the algorithm.

Public research offers a pathway to correcting for systemic bias. For example, in 1994 FRS researchers added "African-American, Hispanic, Asian, Indian, Pacific Islander, and Native American participants."[103] Black box algorithms, on the other hand, present significant barriers to detecting and mitigating biases. The same opacity that makes it difficult for customers and researchers to assess the accuracy of a black box algorithm, or even whether an "error" has occurred, also makes it difficult to assess whether the algorithm's biases reflect societal biases in the way they operate.

Finally, the propensity of algorithms to reflect societal biases can be exacerbated by the propensity for organizations to implement algorithms for unvalidated purposes. If the intended purpose of a model, for example, is to assess the real-world impact of barriers to medical access, it may be important for the model to reflect how those access barriers may correlate with an individual's membership in a protected class. But if the organization subsequently implements that *same model* in a way that influences its decision-making, that implementation will often transform the algorithm's observations about biases that exist in the world into a recommendation for those biases to be perpetuated.

III. *DISSECTING BIAS* AND REGULATORY CONCERNS

A. Empirical Results and Purported Findings of Racial Bias

Dissecting Bias studied how "one of the largest and most typical examples of a class of commercial risk-prediction tools"[104] assigned risk scores to Black and White patients. Comparing those scores to a

103. *See* FRAMINGHAM HEART STUDY, *supra* note 71, at 2.

104. *See Dissecting Bias*, *supra* note 1.

measure of these patients' chronic conditions, the study concludes that this comparison indicates that the "widely used algorithm . . . exhibit[ed] significant racial bias."[105]

In the study, a hospital sought to identify its patients who suffer from multiple chronic conditions. Rather than use an established clinical instrument to assess multiple chronic conditions,[106] the hospital elected to use an algorithm that is often used to predict future medical expenditures.[107] The hospital used the algorithm to assign risk scores to its patients. Patients who had risk scores in the top three percentile[108] were automatically identified for enrollment in a care management program that would provide additional medical services. This program gave patients "greater attention from trained providers, to help ensure that care [wa]s well coordinated,"[109] including "teams of dedicated nurses, extra primary care appointment slots, and other scarce resources," that "are widely considered effective at improving [healthcare] outcomes."[110] Patients whose risk scores were between the third and fifty-fifth percentiles were referred to their primary care physician "to consider whether they would benefit from program enrollment."[111] It appears that patients falling outside either threshold were ineligible for enrollment in the care management program.

The study sought to assess the impact, if any, that this implementation had on patients who self-identified as "Black" versus those who self-identified as, or were presumed to be, "White."[112] It examined

---

105. *Id.*

106. *E.g.*, Greg Walker, Ph.D., *Care Management Dashboards: Calculation of Risk Scores*, R.I. QUALITY INST. https://www.riqi.org/sites/default/files/2018-08/Care_Management_-_Risk_Score_Calculation_-_Summary_GW.pdf (last visited Apr. 4, 2021).

107. *See Dissecting Bias*, *supra* note 1, at 450 (referencing "the algorithm manufacturer's choice to predict future costs").

108. *Id.* at 448 ("Patients above the 97th percentile are automatically identified for enrollment in the program.").

109. *Id.* at 447.

110. *Id.*

111. *Id.* at 448 ("[Patients] above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.").

112. *Id.* at 447 ("We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who

a sample of approximately 50,000 patients, 43,539 of which were assigned the label "White" and 6,079 of which self-identified as "Black."[113] For data sources, the study used the risk score that the algorithm assigned to each patient[114] and a "rich set of outcome data" about the patients.[115] Researchers then created a "comorbidity score" based on the number of "active chronic conditions."[116] This comorbidity score was intended to measure "how many chronic conditions are flaring up . . . not simply an indicator of previously diagnosed chronic conditions."[117] The researchers then compared the patients' risk scores to their comorbidity scores and evaluated how each score would have prioritized patients for eligibility into the care management program described above.

1. Impact on "High Risk" Patients

Based on these comparisons, *Dissecting Bias* concluded that the algorithm's risk scores exhibited "significant racial bias."[118] At a given risk score assigned by the algorithm, the study found that "Black patients are considerably sicker than White patients, as evidenced by

---

self-identified as races other than White (*e.g.*, Hispanic) were so considered. . . . We considered all remaining patients to be White.").

113. *Id.* at 448 ("Our main sample thus consisted of (i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity, whom we observed over 11,929 and 88,080 patient-years, respectively ([One] patient-year represents data collected for an individual patient in a calendar year). The sample was 71.2% enrolled in commercial insurance and 28.8% in Medicare; on average, 50.9 years old; and 63% female.").

114. *Id.* at 448 ("[W]e obtained algorithmic risk scores generated for each patient year. In the health system we studied, risk scores are generated for each patient during the enrollment period for the system's care management program.").

115. *Id.* at 447.

116. *Id.* at 448.

117. Ziad Obermeyer et al., *Supplementary Materials for Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, SCI. MAG., Oct. 25, 2019, at 6, https://science.sciencemag.org/highwire/filestream/733420/field_highwire_adjunct_files/0/aax2342_Obermeyer_SM.pdf ("Of note, this is a measure of how many chronic conditions are flaring up and driving utilization, not simply an indicator of previously diagnosed chronic conditions (for which predictions are not necessarily required).").

118. *See Dissecting Bias*, *supra* note 1.

signs of uncontrolled illnesses."[119] This was indicated, in part, by the fact that Black patients receiving the same "risk score" had "26.3% more chronic illnesses than Whites."[120] This disparity, in turn, impacted the proportion of Black patients identified as "high risk." Using the algorithm's risk scores as a measure of severity, only 17.7% of patients in the highest risk category were Black; however, if the study's comorbidity score was used, that percentage would rise to 46.5%.[121]

2. Impact on "Medium Risk" and Excluded Patients

One potential safeguard against algorithmic bias is having an algorithm's outputs filtered by humans. COMPAS risk scores, for example, are not to be slavishly followed in sentencing a convicted inmate. The same was true for patients at or above the fifty-fifth percentile of risk. According to *Dissecting Bias*, those patients were referred to their primary care physician who made the enrollment determination.[122] These doctors are "presented with contextual information from patients' electronic health records and insurance claims and are prompted to consider enrolling them in the program."[123]

Dissecting Bias attempted to assess the extent to which this human intervention offset the algorithm's observed "significant racial bias." It did this by comparing the racial composition of sample patients who were enrolled in the care management program to simulations of what would be predicted by the implemented algorithm and comorbidity score, respectively.[124] In this experiment, the percentage

119. *Id.*

120. *Id.* at 448.

121. *Id.* at 449 ("At α = 97th percentile, among those auto-identified for the program, the fraction of Black patients would rise from 17.7 to 46.5%.").

122. *Id.* at 448 ("[Patients] above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.").

123. *Id.* at 452 ("Specifically, for patients at or above a certain level of predicted risk (the 55th percentile), doctors are presented with contextual information from patients' electronic health records and insurance claims and are prompted to consider enrolling them in the program. Thus, realized enrollment decisions largely reflect how doctors respond to algorithmic predictions, along with other administrative factors related to eligibility. . . .").

124. *Id.*

of Black enrolled patients would be 18.3%.[125] This was lower than the percentage of Black patients actually enrolled by physicians, which was 19.2%.[126] The percentage of Black patients who "would have been enrolled" based solely on the severity of their comorbidity scores, however, would be 26.9%, a significantly higher percentage.[127] From this, the authors conclude that "although doctors do redress a small part of the algorithm's bias, they do so far less than" using an algorithm based on the patients' medical conditions.[128]

Given the reliance on counterfactual simulations, one must approach such results with caution. The simulation, however, is more probative than simply accepting, as an article of faith, that inserting a human into a process is always an effective safeguard against the systemic influence of an algorithm. As noted in *How Context Affects Choice*, "it is now widely accepted that choices are susceptible to contextual influences."[129] It is possible that physicians may have presumed that the ranking reflected in the patients' risk scores accurately reflected the comparative severity of the patients' underlying medical conditions.

The implementation describes no safeguards for patients with severe multiple chronic conditions who, for whatever reason, were not identified as being in the top fifty-fifth percentile of risk based on their medical expenditure risk scores.[130] Physicians may be unlikely to enroll patients who they believe have already been ruled ineligible based on the hospital's algorithm. This factor alone could make a significant contribution to biased outcomes even in situations where individual physicians exhibit no bias in their medical decision-making. The

---

125. *Id.*

126. *Id.*

127. *Id.*

128. *Id.* at 452 ("Thus, although doctors do redress a small part of the algorithm's bias, they do so far less than an algorithm trained on a different label.").

129. Raphael Thomadsen et al., *How Context Affects Choice*, 5 CUSTOMER NEEDS AND SOLUTIONS 3 (Nov. 25, 2017), https://www.hbs.edu/ris/Publication%20Files/Thomadsen%20et%20al%202017%20-%20How%20Context%20Affects%20Choice_3053d369-3473-45fc-94a3-b18cd100d5d3.pdf.

130. *See Dissecting* Bias, supra note 1, at 448 (discussing that "[p]atients above the 97th percentile are automatically identified for enrollment in the program. Those above the 55th percentile are referred to their primary care physician . . . .", but otherwise making no reference to what occurred with patients scoring at or below the 55th percentile).

algorithm (or its implementation) already did it for them by inaccurately forecasting their medical costs as too low.

B. Regulatory Concerns: Perpetuating Racial Bias and Conflicts of Interest

The response from New York regulators to *Dissecting Bias*'s findings focused on two concerns. The first was the racial bias reported by the study.[131] In response, the regulators called on the manufacturer to cease using the algorithm unless it can "demonstrate that [its] algorithm is not racially discriminatory."[132] Second, the regulators were "troubled by the potential for conflicts of interest to the extent that the entity that controls the algorithm also is affiliated with both providers and insurers, and the algorithm considers costs."[133]

Insurance regulators are familiar with tactics used to deny valid medical claims. In 2017, for example, New York announced a settlement with Oxford Health for improperly denying claims of hundreds of patients needing infusion services.[134] In another example, California's Department of Managed Health Care ("DMHC") settled with Aetna Health of California to stop denying payment for claims under California's standard for emergency room services.[135]

---

131. *See* New York Letter ("We write in response to reports that Optum's data analytics program, Impact Pro, significantly underestimates health needs for [B]lack patients. Specifically, a recent study published in the journal *Science* showed that Impact Pro's flawed algorithm ranked healthier white patients as equally at risk for future health problems—and therefore in need of more intensive healthcare intervention—as [B]lack patients who suffered from far more chronic illnesses.").

132. *Id.*

133. *Id.*

134. *See* Press Release, N.Y. Att'y Gen. Eric Schneiderman, A.G. Schneiderman Announces Settlement with Oxford Health for Improperly Denying Claims of Hundreds of Members for Infusion Services (Feb. 23, 2017), https://ag.ny.gov/press-release/2017/ag-schneiderman-announces-settlement-oxford-health-improperly-denying-claims (announcing settlement with Oxford Health Insurance, Inc. "requiring that Oxford provide refunds to hundreds of small group plan members in New York State for improperly denying coverage of infusion services . . . .").

135. *See, e.g.*, Press Release, Cal. Dep't of Managed Health Care, DMHC Orders Aetna to Stop Wrongfully Denying Payment for Emergency Medical Services (Aug. 25, 2020), http://www.dmhc.ca.gov/AbouttheDMHC/Newsroom/August25,2020.aspx (announcing settlement with Aetna Health of California, Inc., to "stop using the plan's national standard to deny payment for emergency room claims

Regulators are also familiar with subtler tactics where insurers or their sub-contractors throw barriers at patients with the hopes of using those barriers as cost savings tools. California's DMHC, for example, announced an enforcement action against Employee Health Systems Medical ("EHS") and twelve health plans who utilized EHS's services to deliver primary care to Medi-Cal managed care enrollees.[136] California's DMHC found that EHS and its subcontractor, SynerMed, sought to maximize profit by restricting enrollee's access to health services. This was accomplished by restricting patients' access to more expensive health care providers by "secretly concealing costly specialists from EHS's network."[137] The report notes that "[w]hen doctors tried to refer their patients to one of these suppressed doctors, they found that they could choose only from a smaller list of doctors instead of the full list of doctors that should have been available."[138]

This goal of diverting patients to lower-cost care often resulted in extensive delays in authorizing necessary treatment. In one instance, for example, "the authorization involving a kidney transplant was delayed . . . over 100 days."[139] These strategies are distinct from outright denials but can be effective at accomplishing the same goal.

An algorithm "deployed nationwide" could accomplish similar objectives. Insurers seeking to reduce access to preventative medical care could achieve that objective by encouraging providers to use an algorithm that de-prioritizes reaching out to patients who suffer from access barriers.[140] Similarly, when such algorithms are used to calculate payment targets for providers participating in shared savings or other value-based payment models, health care providers are financially incentivized to place the financial interests of the insurer over the

---

[which] resulted in Aetna wrongfully denying members' emergency room claims as the plan should be applying California's broader standard to approve emergency room services.").

136. *See* Press Release, Cal. Dep't of Managed Health Care, DMHC Fines 12 Health Plans $1.9 Million for Improperly Denying Care to Enrollees (Dec. 18, 2019), https://www.dmhc.ca.gov/AbouttheDMHC/Newsroom/December18,2019.aspx.

137. *Id.*

138. *Id.*

139. *Id.*

140. *See id.* (noting this phenomenon with respect to Black patients in that "algorithm appears to inherently prioritize [W]hite patients who have had *greater access* to healthcare than [B]lack patients." (emphasis added)).

healthcare needs of their patients. Rather than driving cost savings through improved medical care, providers are financially dis-incentivized to treat historically disadvantaged patients because those patients' historical costs have been incorporated into payment targets. If providers believe that this same algorithm is capable of making objective assessments of their patients' medical needs, they may be blind to the possibility that their medical decision-making is potentially being influenced by non-medical considerations. They may not, therefore, recognize the need to deploy safeguards to mitigate the impact of such influence.

IV. DIAGNOSING THE PROBLEM

A. Dissecting Bias*'s Diagnosis: Flaw in Algorithmic Design*

Dissecting Bias interprets its findings as showing that the algorithm in question "exhibits significant racial bias."[141] This, in turn, suggests that the flaw is inherent to the algorithm's operation and arises from its manufacture. Recalling the five mechanisms described in *Big Data's Disparate Impact*,[142] *Dissecting Bias* attributes the bias to a problem in "label choice"[143]:

> The dilemma of which label to choose relates to a growing literature on "problem formulation" in data science: the task of turning an often amorphous concept we wish to predict into a concrete variable that can be predicted in a given dataset. Problems in health seem particularly challenging: Health is, by nature, holistic and multidimensional, and there is no single, precise way to measure it. Health care costs, though well measured and readily available in insurance claims data, are also the result of a complex aggregation process with a number of distortions due to structural inequality, incentives, and inefficiency. So although the choice of label is perhaps the single most important decision made in the development

141. *See Dissecting Bias*, *supra* note 1, at 447.

142. *See generally* Barocas & Selbst, *supra* note 2, at 672.

143. This is equivalent to what *Big Data's Disparate Impact* referred to as "target variable." *See* discussion in Section II.E.

> of a prediction algorithm, in our setting and in many others, there is often a confusingly large array of different options, each with its own profile of costs and benefits.[144]

According to *Dissecting Bias*, manufacturer's problem in developing its algorithm was selecting the wrong label. Rather than select the label "total costs" as the algorithm's target variable, the manufacturer should have used "active chronic conditions."[145]

B. Dissecting Bias*'s Assumption and its Impact on Identifying the Source of Bias*

Dissecting Bias's interpretation—that the algorithm "exhibits significant racial bias"—is premised on a key assumption: that the algorithm was used for a purpose that it was designed for, and that its outputs had been empirically validated to support the selection of medical conditions. There are reasons to question this assumption.

Dissecting Bias acknowledges, for example, that the "bias arises because the algorithm predicts health care *costs* rather than illness."[146] This was confirmed by the algorithm's vendor, who stated "the cost model within Impact Pro was highly predictive of cost, which is what it was designed to do."[147] Moreover, many of the algorithm's predictive capabilities are described by the SOA Report. However, *none* of the SOA Report's assessments discuss the algorithm's ability to predict a patient's medical conditions.

Notably, when the algorithm's outputs were used solely for cost-prediction purposes, the researchers observed *no* evidence of bias:

> As a first check on this potential mechanism of bias, we calculate the distribution of realized costs *C* versus predicted costs *R*. By this metric, one could call the algorithm *unbiased*. . . . [A]t every level of algorithm-

144. *See Dissecting Bias*, *supra* note 1, at 451.

145. *Id.* ("Alternatively, rather than predicting costs at all, we could simply predict a measure of health; e.g., the number of active chronic health conditions.").

146. *Id.* (emphasis added).

147. Christopher Snowbeck, *Regulators Probe Racial Bias with UnitedHealth Algorithm*, STAR TRIB.(Oct. 28, 2019 6:59 PM), https://www.startribune.com/regulators-probe-racial-bias-with-unitedhealth-algorithm/563997722/ (quoting a UnitedHealth statement).

> predicted risk, Blacks and Whites have (roughly) the same costs the following year. . . .Conditional on risk score, predictions do not favor Whites or Blacks anywhere in the risk distribution.[148]

If so, this suggests that the "significant racial bias" is *not* a property of the algorithm itself. Rather, it arose from *how* the algorithm and its risk scores were used to identify patients who had multiple chronic conditions.

However, this does not absolve the manufacturer of responsibility. As *Dissecting Bias* notes in its *Supplementary Materials*, the manufacturer's marketing materials appear to tout the algorithm's ability to identify individual patients in need of medical attention:

> The algorithm's stated goal (from promotional materials) is to predict which individuals are in need of specialized intervention programs and which intervention programs have the most impact on the quality of individuals' health. These scores, which are meant to flag individuals for intervention before their health becomes catastrophic, are a key part of the decision to enroll a patient in the care management program . . . .[149]

Further, this does not mute the gravity of *Dissecting Bias*'s results. Evidence of "significant racial bias" is alarming regardless of its source, particularly if it is "typical of [an] industry-wide approach."[150]

However, it does suggest that the source of the bias is *not* something that can be fixed by changing the algorithm's label choice.[151] Rather, that it can only be addressed through greater attention to how decision-making algorithms are implemented.

As previously noted, "a risk scoring model designed for one outcome . . . may not be a suitable risk scoring model for another outcome . . . ."[152] As discussed in Section II.B above, the accuracy of an algorithm's predictions can vary significantly depending on the type of

148. *See Dissecting Bias*, *supra* note 1, at 449–50.

149. Obermeyer et al., *supra* note 117, at 3.

150. *See Dissecting Bias*, *supra* note 1.

151. *Id.* at 452 ("Bias attributable to label choice . . . is a useful framework through which to understand bias in algorithms . . . .").

152. HILEMAN ET AL., *supra* note 29.

question it is asked to answer. Moreover, the use of an algorithm designed for one purpose can often have adverse consequences when the same algorithm is used for a secondary purpose. One of those consequences can be to convert observed disparate impacts into recommendations to perpetuate those biases.

Here, it appears that the "significant racial bias" arose from the decision to use a cost-prediction algorithm to identify individuals with multiple chronic conditions. We must, therefore, turn to the inherent risks of using a cost prediction algorithm to influence medical decisions, and evidence that the vendor and its customer understood and mitigated those risks.

C. Reasonably Foreseeable Risks When Using Cost Algorithms to Predict Medical Conditions

There are many ways that a hospital could identify patients who may be suffering multiple chronic conditions. If a hospital is attempting to assess all of its patients with technology, it could utilize an established measure of chronic condition comorbidity[153] using its patients' electronic health record ("EHR") information. Alternatively, it could develop its own "comorbidity score" to measure "active chronic conditions" along the lines of the researchers in *Dissecting Bias*.[154]

Rather than taking this direct approach, however, the hospital used a cost-prediction algorithm as an indirect means of identifying patients who have multiple chronic conditions.[155] Moreover, even though a cost prediction algorithm is incapable of directly indicating whether a patient has multiple chronic conditions, it appears that the hospital used the comparative magnitude of the patients' risk scores as a proxy for their likelihood of having multiple chronic conditions. As noted by the study's authors, "[p]atients above the 97th percentile are

153. Greg Walker, Ph.D., *Care Management Dashboards: Calculation of Risk Scores*, R.I. QUALITY INST. https://www.riqi.org/sites/default/files/2018-08/Care_Management_-_Risk_Score_Calculation_-_Summary_GW.pdf (last visited Apr. 4, 2021).

154. *See Dissecting Bias*, *supra* note 1, at 448 ("We begin by calculating an overall measure of health status, the number of active chronic conditions (or 'comorbidity score,' . . . ")).

155. *Id.* ("In the health system we studied, risk scores are generated for each patient during the enrollment period for the system's care management program.").

automatically identified for enrollment in the program. Those above the 55th percentile [were] referred to their primary care physician, ..."[156]

1. Accuracy Risks

i. Accuracy of Comparative Cost Predictions Versus Direct Comorbidity Measures

According to *Dissecting Bias*, the hospital selected patients for the enhanced care coordination services by comparing their cost risk scores to one another.[157] Patients who had risk scores in the top 3 percentile were automatically identified for enrollment in the care management program. Patients whose risk scores were between the 3rd and 55th percentile were referred to their primary care physician "to consider whether [the patients] would benefit from program enrollment."[158] It appears that patients falling outside either threshold were ineligible to receive the care coordination program offered to "high" and the "moderate" risk patients.[159]

Although *Dissecting Bias* provides insufficient detail to answer definitively, it seems unlikely that the hospital intended to rely on the algorithm to conclusively determine whether or not its patients had multiple chronic conditions. Rather, as discussed in Section D.2 below, it appears more likely that the hospital intended for the algorithm to preliminarily identify which of its patients *may* have multiple chronic conditions. It is possible that the hospital chose this approach out of convenience in light of the potential time and expense associated with extracting data from its EHR data warehouse.[160] Even if this were the case, however, the use of a cost-prediction algorithm in a way that

156. *Id.*

157. *See* discussion *supra* Section IV.A.

158. *See Dissecting Bias*, *supra* note 1, at 448.

159. *Id.*

160. Obermeyer et al., *supra* note 117, at 5 ("Of note, these [electronic health record] data are not routinely analyzable, as they must be pulled and cleaned extensively from hospital data warehouses. As a result, algorithm developers typically do not have access to them to fit or validate predictions, making this exercise particularly useful to assess algorithm performance in general.").

could influence clinical decision-making introduces inherent risks to patient care that must be understood if they are to be mitigated.[161]

ii. Data Quality

Among the many benefits of using a clinically validated measure of chronic comorbidity is that it utilizes clinically relevant information directly from the patients' electronic health records. This is, in fact, what the researchers utilized in *Dissecting Bias* in their comorbidity score: "[t]o measure [patients' health], we link predictions to a wide range of outcomes in *electronic health record data*, including all diagnoses . . . as well as key quantitative laboratory studies and vital signs capturing the severity of chronic illnesses."[162]

As noted in *Dissecting Bias*, "algorithm developers typically do not have access to [EHR data] to fit or validate their predictions."[163] It is surprising, therefore, if the comparative deficit in health information about the patients adversely impacts the accuracy of their predictions. On the face of it, it is predictable that the comparative richness of hospital's data would be a superior source from which to make accurate predictions about the hospital's patients.

iii. Using Comparative Magnitude of Predicted Costs as a Proxy for Multiple Chronic Conditions

One problem with measuring a patient's propensity for having a chronic condition based on how her predicted medical costs compare to those of other patients are the error ranges of all of the predictions involved. In the SOA Report, for example, one algorithm's predictions about individuals is correct 43.4% of the time when its error tolerance is set to plus or minus 25% of predicted costs.[164] When comparing a patient with a risk score of, *e.g.*, 1, to a second patient with a risk score of 1.14, the hospital would not know which of those patients falls within the "tolerable error" range. Moreover, even within the error

161. Thomadsen et al., *supra* note 129 ("[I]t is now widely accepted that choices are susceptible to contextual influences.").

162. *See Dissecting Bias*, *supra* note 1, at 448 (emphasis added).

163. *See* Obermeyer et al., *supra* note 117, at 5.

164. *See* HILEMAN & STEELE, *supra* note 35, at 40 (reviewing column for "CDPS" in Table 4.5.1)

range, the hospital does not know which of those risk scores are "correct." The hospital, therefore, can place only limited confidence on the likelihood that the 1.14 risk score is indicative of likely higher costs than the 1 score, and it can be even less confident that either score reflects the comparative severity of the patients' respective medical conditions.

A major exception to this general rule is the risk scores of patients predicted to be in the "top 1%" of medical costs. As discussed, the SOA Report indicated that *all* of the algorithms' ability to identify the "top 1%" of most costly patients had AUCs of 0.8 to 0.9,[165] regarded as "excellent" in assessing an instrument's predictive ability.[166] This accuracy, however, does not translate to the "other 99%." As previously discussed, even when the tolerable error rate is set to a maximum of 1.0, the algorithms were only measured to have AUCs of approximately 0.5.[167]

2. Risk of Biased Recommendations Arising from Historical Cost Data

Yet another challenge of utilizing a cost-prediction algorithm to inform medical decisions is the extent to which the algorithm's risk scores are biased by the cost data it relies on. As noted in *Big Data's Disparate Impact*, one mechanism giving rise to biased outcomes is where the algorithm "draws inferences from a biased sample of the population."[168] When cost risk scores are used for the purposes of calculating near term transfer payments, any bias in the data may have no influence over healthcare decision-making. When, however, that same algorithm is used to influence medical decisions about specific individuals, "any decision that rests on these inferences may systematically disadvantage those who are under–or overrepresented in the dataset."[169]

The types of patients likely to be under-represented in historical cost information are those who have fewer medical costs *because* they have trouble accessing medical care. *Inequality in Quality, Addressing*

165. *Id.* at 46.

166. *See* Mandrekar, *supra* note 54, at 1316.

167. HILEMAN & STEELE, *supra* note 35, at 43.

168. *See* Barocas & Selbst, *supra* note 2, at 681.

169. *Id.*

Socioeconomic, Racial, and Ethnic Disparities in Health Care[170] ("*Inequality in Quality*"), for example, notes that "[s]ocioeconomic and racial/ethnic disparities in health care quality have been extensively documented." Citing dozens of studies, *Inequality in Quality* summarizes their results as follows:

> In the United States, lower socioeconomic position is associated with lower overall health care use, even among those with health insurance. Socioeconomic position, as measured by education or income, is also clearly related to standard measures of health care quality. Lower socioeconomic position is associated with receiving fewer Papanicolaou tests, mammograms, childhood and influenza immunizations, and diabetic eye examinations, later enrollment in prenatal care, and lower quality ambulatory and hospital care.[171]

Moreover, those barriers disproportionately impact members of specific racial and ethnic backgrounds:

> Elderly [B]lacks, compared with [W]hites, are seen less often by specialists, receive less appropriate preventive care including mammography and influenza vaccinations, lower quality hospital care, and fewer expensive, technological procedures. In general, [B]lacks receive less intensive hospital care, including fewer cardiovascular procedures, lung resections for cancer, kidney and bone marrow transplants, cesarean sections, peripheral vascular procedures, and orthopedic procedures. They have also been reported to receive less aggressive treatment of prostate cancer, fewer antiretrovirals for human immunodeficiency virus infection, antidepressants for

170. Kevin Fiscella et al., *Inequality in Quality: Addressing Socioeconomic, Racial, and Ethnic Disparities in Health Care*, 283 J. AM.MED. ASS'N 2579 (2000) [hereinafter *Inequality in Quality*], https://jamanetwork.com/journals/jama/article-abstract/192714.

171. *Id.* (citing approximately a dozen studies).

> depression, tympanostomy tubes, and admissions for chest pain, and lower-quality prenatal care.[172]
>
> Other racial and ethnic groups are also impacted:
>
> Compared with Whites, Latinas receive fewer mammograms, Papanicolaou tests, and influenza vaccinations, less prenatal care, fewer cardiovascular procedures, and less analgesia for metastatic cancer and trauma. Asian Americans receive fewer Papanicolaou tests and influenza vaccinations. Native Americans receive less prenatal care.[173]

These widely documented access barriers, in turn, are reflected in the historical cost information pertaining to those patients.[174] It seems highly likely, therefore, that any algorithm that bases its predictions on historical costs will underestimate the severity of the medical needs of patients who have experienced medical access barriers.[175] It is this phenomenon, in fact, that appears to be confirmed by *Dissecting Bias*'s results. Moreover, the algorithm's vendor appeared to acknowledge this reality, when it responded to *Dissecting Bias* by noting that "[t]hese gaps [are] often caused by social determinants of care and other socio-economic factors."[176]

3. Inherent Challenges Validating a "Black Box" Algorithm

Compounding the risks described above is the fact that the algorithm in question appears to be a "black box" algorithm where the manufacturer does not disclose the relationship between the algorithm's inputs and outputs. This places meaningful barriers on a customer's ability to empirically evaluate the inherent risk of using the algorithm

172. *Id.* at 2579–80 (citing approximately two dozen studies).

173. *Id.* at 2580 (citing a half-dozen studies).

174. *See Dissecting Bias*, *supra* note 1, at 450 ("These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities.").

175. *Id.* ("As a result, accurate prediction of costs necessarily means being racially biased on health.").

176. Snowbeck, *supra* note 147.

and insulating its patients from those risks. *Dissecting Bias* described its experience navigating these barriers as follows:

> Empirical investigations of algorithmic bias . . . have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work 'from the outside,' often with great ingenuity, and resort to clever work arounds such as audit studies.[177]

Customers are often in no better position than researchers in being able to empirically verify the accuracy and suitability of an algorithm being licensed. Audit studies, such as the one performed in *Dissecting Bias* or the ones performed by various states with respect to COMPAS risk scores,[178] can validate some aspects of an algorithm. However, even when problems are identified in an audit study, these barriers often make it impossible to identify the source of the problem, much less "figur[e] out what to do about them."[179]

4. Need for Proactive Management of Foreseeable Risks

In summary, the implementation described in *Dissecting Bias* presents a number of significant risks. First, the hospital elected to not identify patients using an established clinical measure of comorbidity

177. *See Dissecting Bias*, *supra* note 1.

178. *See, e.g.*, Sharon Lansing, N.Y. State Div. of Crim. Just. Servs., Off. of Just. Rsch. and Performance, *New York State COMPAS Probation Risk and Need Assessment Study: Examining the Recidivism Scale's Effectiveness and Predictive Accuracy* (Sept. 2012), http://www.northpointeinc.com/downloads/research/DCJS_OPCA_COMPAS_Probation_Validity.pdf; Jennifer L. Skeem & Jennifer E. Loudon, *Report Prepared for the California Department of Corrections and Rehabilitation (CDCR), Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)* (2007) [hereinafter *California COMPAS Study*], http://risk-resilience.berkeley.edu/journal-article/assessment-evidence-quality-correctional-offender-management-profiling-alternative.

179. *See Dissecting Bias*, *supra* note 1 ("[Audit studies] can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves.").

utilizing its own EHR data.[180] Instead, *Dissecting Bias* suggests that the hospital anchored its identification of patients on an algorithm designed to make actuarial cost predictions about groups of patients.[181] According to public benchmarks, the algorithm's ability to make predictions about individuals is limited to predicting medical costs; and even there it is subject to error tolerance thresholds that can frustrate the effective use of its predictions to accurately stratify patients into meaningful clinical groupings within those tolerances.[182] Moreover, the cost data that the algorithm uses is likely of significantly lower quality than the EHR data the hospital already possesses; and the algorithm's data is likely biased in a manner that, among other things, reflects societal biases.[183]

Absent proactive mitigation measures, the implementation described in *Dissecting Bias* appears highly susceptible to adverse consequences, from excluding patients in urgent need of medical care to perpetuating racially biased outcomes. The questions to consider next, therefore, is what mitigation steps the hospital may have taken to insulate its patients from these risks.

D. Mitigation of Foreseeable Risks

Dissecting Bias does not directly discuss any safeguards the hospital may have deployed to address the numerous prima facie risks discussed above.[184] Certain aspects of the way the hospital implemented the algorithm, however, offer clues as to steps the hospital may have taken. To understand the options available to the hospital, we will first look at a "chronic care management" ("CCM") program, launched by CMS, which sought to offer the same kind of care management benefits to the same kind of patients suffering from multiple chronic conditions.

180. *See supra* discussion Section III.A.

181. *See* discussion *supra* Section III.A; *see also* discussion *supra* Section II.B.

182. *See* discussion *supra* Section III.A; *see also* discussion *supra* Section II.B.

183. *See supra* discussion Section IV.C.

184. *See supra* discussion Section IV.C.

1. CMS's Chronic Care Management Program: Insulating Treatment Decisions

CMS's CCM program sought to give patients suffering from multiple chronic conditions enhanced care management services. To be eligible, a patient must have "multiple chronic conditions expected to last at least twelve months or until the death of the patient, and that place the patient at significant risk of death, acute exacerbation/decompensation, or functional decline."[185] Under the program, patients would receive a variety of care coordination benefits, including, among other things, "24/7 access to physicians or other qualified health care professionals"; "[c]omprehensive care management . . . to ensure timely receipt of all recommended preventive care services"; "[m]anagement of care transitions between and among health care providers and settings"; and "[c]oordination with home- and community-based clinical service providers."[186]

CMS had a financial rationale for launching CCM. It sought to reduce the long-term healthcare costs for patients with multiple chronic conditions who were at high risk of suffering significant catastrophic medical events.[187] The pathway to achieving those savings was to provide proactive care coordination services to those patients.

185. Medicare Program; Revisions to Payment Policies Under the Physician Fee Schedule, Clinical Laboratory Fee Schedule & Other Revisions to Part B for CY 2014, 78 Fed. Reg. 73993, 74415 (Dec. 10, 2013) [hereinafter 2013 CCM Final Rule], https://www.govinfo.gov/content/pkg/FR-2013-12-10/pdf/FR-2013-12-10.pdf.

186. Medicare Program; Revisions to Payment Policies Under the Physician Fee Schedule and Other Revisions to Part B for CY 2018; Medicare Shared Savings Program Requirements; and Medicare Diabetes Prevention Program, 82 Fed. Reg. 52976, 53174 (Nov. 15, 2017) [hereinafter 2017 Rule], https://www.govinfo.gov/content/pkg/FR-2017-11-15/pdf/2017-23953.pdf.

187. *See, e.g.*, JOHN SCHURRER, ET AL., MATHEMATICA POLICY RESEARCH, EVALUATION OF THE DIFFUSION AND IMPACT OF THE CHRONIC CARE MANAGEMENT (CCM) SERVICES: FINAL REPORT 58 (Nov. 2, 2017), https://innovation.cms.gov/files/reports/chronic-care-mngmt-finalevalrpt.pdf (report prepared on behalf of CMS) ("[W]e calculated gross savings to the Medicare program associated with CCM over a 12-month period following first receipt of CCM services as $88 million.")

"Beneficiaries receive high quality, coordinated, effective, efficient care. [And] [a]s a result, health care costs are reduced."[188]

CMS's implementation of CCM avoids all of the risks described above by isolating treatment decisions from individual patient's "risk scores." Eligibility for the program, for example, does not take into consideration the patient's historical medical charges, nor an assessment of the severity of her medical conditions compared to other patients.[189] Consequently, patients who have experienced access barriers will not be deemed less eligible for CCM benefits simply because of their history of medical claims. Moreover, eligibility is extended to *all* patients that satisfy CCM's criteria regardless of the severity of another patient's condition. CMS's eligibility criteria, for example, does not presume that no more than 45% of a physician's patients are eligible for the program.[190]

CMS's CCM program wholly insulates individual patients from the impact of whatever actuarial cost information CMS utilized to assess the financial impact of the CCM program or design CCM's program dimensions.

2. Risk Mitigation the Hospital May Have Deployed

Dissecting Bias offers little detail about the hospital's implementation. It is unclear, for example, whether the hospital conducted a data impact assessment of the algorithm that considered the risks identified above. It is likewise unclear what safeguards the hospital may have implemented to address those risks. Accordingly, any discussion regarding the hypothetical steps the hospital may have taken is

188. *See, e.g.*, CTRS. FOR MEDICARE AND MEDICAID SERVS., CMS STRATEGY: THE ROAD FORWARD 2013-2017 3 (2013), https://www.nadona.org/wp-content/uploads/2016/05/CMS-Strategy.pdf.

189. *See, e.g.*, 2017 Rule, *supra* note 186, at 53170 (Patient eligibility does not require billing provider to conduct comparisons among her patients in determining whether or not a patient is eligible for the program. Rather "patient eligibility" is defined as applying to any patient with "multiple (two or more) chronic conditions expected to last at least 12 months or until the death of the patient, and that would place the patient at significant risk of death, acute exacerbation/decompensation, or functional decline . . . . " None of these criteria require or authorize a provider to base eligibility determinations by comparing a patient's medical conditions to other patients.).

190. *Id.*

inherently speculative. As discussed below, however, *Dissecting Bias*'s description offers some details that suggest certain safeguards may have deployed.

As previously discussed, it seems unlikely that the hospital intended to rely on the algorithm to conclusively determine whether or not its patients had multiple chronic conditions. Rather, it appears more likely that the hospital used the algorithm to make preliminary identifications of patients that *may* have multiple chronic conditions. Patients who had risk scores in the top third percentile[191] were automatically identified for enrollment in the hospital's care management program. Patients whose risk scores were between the 3rd and 55th percentile were referred to their primary care physician "to consider whether they would benefit from program enrollment."[192]

As previously discussed, the SOA Report rated all cost-algorithms as highly effective at predicting patients in the "top 1%" of most costly patients,[193] including Impact Pro, which had AUCs ranging between 0.841 and 0.871.[194] The hospital may have concluded that this performance was sufficiently accurate for the purposes of identifying patients whose multiple chronic conditions were severe enough so as generate "top 3%" cost predictions.

The vulnerabilities of this approach are fourfold. First, the SOA Report assessed the algorithm's ability to predict costs, *not* medical conditions. Second, the reported AUC ranges only applied to patients in the "top 1%," *not* the "top 3%." Third, the approach does not protect patients from the algorithm's inherent error rate. Nor does it protect patients who suffer access barriers (including minorities) from being adversely impacted by the algorithm's cost data.

The hospital may have sought to address these risks by assigning the remainder of the "top 45%" risk scores to physicians. The thinking may have been that any patients inappropriately left out of the "high" risk category would have a second opportunity to participate via the professional assessment of their primary care physician. *Dissecting*

---

191. *See Dissecting Bias*, *supra* note 1, at 448 ("Patients above the 97th percentile are automatically identified for enrollment in the program.").

192. *Id.* ("[Patients] above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.").

193. HILEMAN & STEELE, *supra* note 35, at 46 fig.5.

194. *Id.*

Bias does not discuss what may have been the assumed "incident rate" of "chronic comorbidity" among the hospital's patient population. It is possible that the selection of the "top 45%" threshold was intended to be over-inclusive, to address the likelihood of patients suffering from multiple chronic conditions otherwise being misclassified.

If this were the thinking, it is unclear that even a "top 45%" threshold would be sufficient. As previously discussed, the AUCs of algorithms predicting patients outside the "top 1%" was notably lower. In the case of Impact Pro, for example, its AUCs ranged between 48.7% and 52.2% when its tolerable error rate was set to 1.0.[195] The accuracy increased to between 76% and 77.7% when the rate was increased to 3.0.[196] It is unclear whether a "top 45%" buffer would be sufficient to address what appears to be fairly significant variances among the patients' likely scores. If this buffer is insufficient, *Dissecting Bias*'s results indicate that a significant number of patients otherwise qualified to receive care coordination services may have been precluded from receiving those services as a result of an observable bias in the algorithm's outputs.[197]

V. QUESTIONS OF DATA GOVERNANCE RAISED BY *DISSECTING BIAS*

A. Deficits in the Manufacturer's Response

The fallout from *Dissecting Bias*'s publication is notable in how little assurance has been offered to patients. According to the study, its results show that a "widely used algorithm . . . exhibits *significant racial bias*."[198] Further, the study purports to show that the algorithm may be adversely impacting thousands—*if not millions*—of patients.[199]

In response, the manufacturer acknowledged that the data powering its algorithm may reflect the very biases *Dissecting Bias*

195. *Id.* at 43 tbl.4.5.3.

196. *Id.*

197. *See supra* discussion Section III.A.2.

198. *See Dissecting Bias*, *supra* note 1.

199. *Id.* ("We show that a widely used algorithm, typical of this industry-wide approach and *affecting millions of patients*, exhibits significant racial bias. . . . It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are *applied to roughly 200 million people* in the United States each year." (emphasis added)).

describes. Such "gaps," the manufacturer notes, are "often caused by social determinants of care and other socio-economic factors."[200] And the manufacturer does not challenge the study's findings. Instead, the manufacturer delegates responsibility for detecting and mitigating such harms to its customers. "These gaps," the manufacturer notes, "can then be addressed by the health systems and doctors to ensure people, especially in underserved populations, get effective, individualized care."[201] Notably, the manufacturer does not state that it *requires* its customers to address such "gaps," nor even that it advises them to.

This posture might be understandable if the algorithm were being used in an unauthorized manner. If the algorithm were solely marketed as a "cost prediction tool[]," for example, it would be difficult for any manufacturer to anticipate every unauthorized use. The manufacturer's marketing materials, however, suggest differently. In a marketing brochure quoted by *Dissecting Bias* and still available online, the manufacturer markets its algorithms' ability to "[h]elp . . . determine which individuals are in need of specialized intervention programs and which intervention programs have the most impact on the quality of individuals' health," as well as to "[f]lag individuals for intervention before their health becomes catastrophic using [the algorithm's] predictive modeling technology."[203]

The brochure also promotes the algorithm's ability to "identify individuals with upcoming evidence-based medicine gaps in care for proactive engagement" and to "[a]ssess intervention opportunities using clinical guidelines and a comprehensive overview of each individual's health and predicted future risks which present opportunities for early intervention."[204] The brochure also touts the algorithms' ability to help population health management teams identify and stratify populations by predicting future risk; discover new intervention program opportunities for individuals; deliver actionable clinical information directly to physicians and group practice managers; and view customized

200. Snowbeck, *supra* note 147.

201. *Id.*

202. *Id.* (quoting the manufacturer as stating that "the cost model within Impact Pro was highly predictive of cost, which is what it was designed to do").

203. *See, e.g.*, Optum, *Impact Pro for Care Management*, OPTUMINSIGHT (May 12, 2012), https://www.optum.com.br/content/dam/optum/resources/productSheets/Impact_Pro_for_Care_Management_ps_06_2012.pdf.

204. *Id.*

provider or group level evidence-based measure rule management reports.[205]

The brochure does not warn prospective customers about the limitations of the algorithm's predictive accuracy described in SOA Report. There is also no discussion about the algorithm's "gaps . . . caused by social determinants of care and other socio-economic factors."[206] Thus, the brochure does not caution customers of the possibility that they may need to take proactive measures "to ensure people, especially in underserved populations, get effective, individualized care."[207]

At the same time, the brochure recognizes that some customers may want to supplement algorithm's base capabilities with additional data and services. Customers can, for example, "[a]ugment information . . . by employing the optional [clinical data integration services] to capture [EHR] data . . . and augmenting individual analytics with data not available in administrative claims, such as BMI and smoking status."[208] Moreover, the algorithm supports configurations that incorporate standard clinically-validated measures, which involve "500 measures including both retrospective and prospective analysis to identify existing and upcoming evidence-based medicine gaps in care."[209] However, the brochure gives no indication that these enhancements are necessary or advisable to avoid risks that could arise from utilizing the algorithm's base configuration.

This apparent *caveat emptor* approach to data governance offers little assurance to patients who may be adversely impacted by the algorithm's "gaps." It also offers little guidance to customers. If it is up to clients to "ensure people, especially in underserved populations, get effective, individualized care," they would be well served by being notified about any "gaps" the manufacturer is aware of. They also would be well served if the manufacturer developed and publicized best practices on how to implement its algorithm in a manner that insulates patients form the types of harms identified by *Dissecting Bias*. These safeguards would be particularly useful in light of the fact that the

205. *Id.*
206. Snowbeck, *supra* note 147.
207. *Id.*
208. Optum, *supra* note 203.
209. *Id.*

algorithm in question is a "black box," making it all the more difficult for customers to detect or remedy any such "gaps" on their own.

This *caveat emptor* public posture may have played a role in the New York regulators' response. Not only does the posture convey a *blasé* indifference to the adverse impacts the algorithm may be having on "millions of patients," it may have sparked concerns over potential "conflicts of interest."[210] To the extent that the algorithm's "gaps" hinder the delivery of medical care to patients with access barriers, the dearth of public guidance on how to address those gaps may be viewed of as a tolerance, if not approval, of those outcomes.

B. Responsible Management of Algorithmic Unaccountability

The manufacturer's response is a frank reminder to healthcare organizations that they are ultimately responsible for how they use algorithms to inform their medical decision-making. A manufacturer's brochure may claim that its algorithms can "flag individuals for intervention." But that same algorithm may be only "highly predictive of cost" because *that* is what it was "designed to do." The manufacturer's brochure also may not warn customers about "gaps" in its data "caused by social determinants of care and other socio-economic factors." Or that these gaps may even "perpetuate racially disparate impacts." It is ultimately the responsibility of "health systems and doctors to ensure people, especially in underserved populations, get effective, individualized care," even in situations where the manufacturer elects not to inform customers about those risks.

The mitigation strategies described in Section D of Part IV reflect two approaches to managing the inherent risks of using algorithmic data in medical decision-making.[211] Each is grounded in the recognition that the outputs of any algorithm are, ultimately, simply *information*. As such, the potentially harmful effects of decision-making algorithms can be addressed by applying familiar data governance principles to those outputs.

---

210. *See* New York Letter, *supra* note 16.

211. *See supra* discussion Section IV.D.

1. Risk Assessments of Algorithmic Outputs

Under HIPAA's Security Rule,[212] a healthcare organization[213] is required to "[e]nsure the confidentiality, integrity, and availability of all [patient information that it] creates, receives, maintains, or transmits."[214] Here, the term "integrity" usually pertains to assurances that patient information is not altered or destroyed in an unauthorized manner.[215] If, however, that term were expanded to cover the *reliability* of the outputs of decision-making algorithms, the Security Rule's security management process would be an effective tool for managing the algorithm's inherent risks.

Under Section 308(a)(1)(ii)(A) of the Security Rule, for example, healthcare organizations would be required to conduct an accurate and thorough assessment of the potential risks and vulnerabilities to the reliability algorithmic data outputs.[216] This accurate and thorough assessment would apply to "any reasonably anticipated threats or hazards

212. 45 C.F.R. §§ 160.101–552, 164.102–106, 164.302–.318 (2013).

213. HIPAA's provisions apply to organizations that it calls "covered entities" and "business associates." Covered entities includes a wide range of "health care providers," including hospitals, medical practices, pharmacies, clinical labs and many others; to health insurers, which HIPAA calls "health plans;" and to entities that process medical claims, known as "health care clearinghouses." The term "business associate" applies to organizations that provide services to covered entities that require the organization to receive access to identifiable information about the covered entities' patients or beneficiaries. See definitions for "covered entity," "health care provider," "health plan," "health care clearinghouse," and "business associate" in 45 C.F.R. § 160.103 (2014). For brevity, this Article refers to covered entities and business associates collectively as "healthcare organizations."

214. 45 C.F.R. § 164.306(a)(1) (2013) ("Covered entities and business associates must . . . [e]nsure the confidentiality, integrity, and availability of all electronic protected health information the covered entity or business associate creates, receives, maintains, or transmits.").

215. *See* 45 C.F.R. § 164.304 (2013) ("*Integrity* means the property that data or information have not been altered or destroyed in an unauthorized manner.") (emphasis in original)).

216. *See* 45 C.F.R. § 164.308(a)(1)(ii)(A) (2013) ("A covered entity or business associate must . . . [c]onduct an accurate and thorough assessment of the potential risks and vulnerabilities to the . . . integrity . . . of electronic protected health information held by the covered entity or business associate.").

to [the reliability of the outputs]."[217] Three main categories of "reasonably anticipated threats and hazards" are the following:

i. Accuracy Risks: The accuracy of an algorithm's predictions is very sensitive to what specific variables the algorithm is being asked to predict. The same algorithm may accurately predict the medical costs for one demographic category but be very inaccurate in others. And it may be very inaccurate, or have no predictive utility, in its predictions about individuals absent the benefit of large error tolerances.[218] Thus, even slight changes to an implementation can result in a significant degradation to the accuracy of the algorithm's *outputs*. Moreover, consistent with the guidance of the International Actuarial Association, all users of decision-making algorithms understand the situations in which an algorithm's results would be unreliable, namely:
   a) The data are insufficiently representative of the underlying situation;
   b) The implicit assumptions that drive the models accuracy are no longer valid; and
   c) The explicit assumptions are no longer valid because the environment is not sufficiently similar to the situation when the assumptions were formed.[219]

ii. Propensity to Reflect Societal Biases: *Big Data's Disparate Impact* describes five mechanisms that result in the propensity for algorithms to reflect societal biases whenever the algorithms rely on real world data in their development or operation. This inherent risk cannot be mitigated absent

217. 45 C.F.R. § 164.306(a)(2) (2013) ("Covered entities and business associates must . . . [p]rotect against any reasonably anticipated threats or hazards to . . . integrity of such information.").

218. *See* discussion *supra* Section II.B.

219. INT'L ACTUARIAL ASS'N, *supra* note 75.

affirmative countermeasures specifically aimed at addressing it.[220]

iii. Propensity for Entities to Mis-Use Algorithmic Outputs: Absent effective safeguards, entities have a demonstrated propensity for using algorithmic *outputs* for purposes that fall outside their confirmed predictive accuracy.[221]

In conducting such assessments, healthcare organizations cannot simply rely on statements made in marketing brochures. If a vulnerability of the algorithm cannot otherwise be mitigated, healthcare organizations should perform their own validation of the algorithms outputs, such as those that were conducted by a number of jurisdictions that implemented COMPAS risk scores.[222] Such assessments are also advisable in situations where the algorithm's predictions are utilized in a manner that falls outside their confirmed predictive accuracy.

2. Mitigating Identified Risks

Following the aforementioned risk analysis, Section 308(a)(1)(ii)(B) of the Security Rule requires healthcare organizations to implement measures sufficient to reduce the identified risks and vulnerabilities to a reasonable and appropriate level.[223] Ideally, all of the performance characteristics of an algorithm will be fully understood. In many situations, however, one or more of an algorithm's capabilities cannot be directly verified. Then effective mitigation will require isolating the algorithmic outputs to environments where their potential inaccuracy or biases will be incapable of adversely impacting patient care, including perpetuating racial or other forms of biased decision-making. CMS's CCM is an example of one such approach,[224] and the

220. *See* discussion *supra* Section II.E.

221. *See* discussion *supra* Section II.C.

222. *See, e.g.*, Lansing, *supra* note 178.

223. *See* 45 C.F.R. § 164.308(a)(1)(ii)(B) (2013) ("A covered entity or business associate must . . . [i]mplement security measures sufficient to reduce risks and vulnerabilities to a reasonable and appropriate level to comply with § 164.306(a)").

224. *See supra* discussion Section IV.D.1.

hypothetical framework described in Section D.2 of Part IV may be another.[225]

In situations where a more elegant solution cannot be engineered, healthcare organizations can consider the Wisconsin Supreme Court's approach to addressing the risks described above.[226] The court placed conditions on the state's use of COMPAS risk scores in sentencing proceeding. One required the state to include the following notice in *every* presentence report that contained COMPAS risk scores so that the adjudicators are advised of the risk scores' limitations and risks:

> The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined. Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders—not a particular high-risk individual. Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism. A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed. Risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations. COMPAS was not developed for use at sentencing, but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole.[227]

The proposed approach would be a relatively minor enhancement of the policies and procedures the majority of healthcare organizations already required to have in place. But they would go a long way to avoid the types of harms described in *Dissecting Bias* and to inspire greater confidence that decision-making algorithms are managed in accordance with the responsibilities they are being assigned.

---

225. *See supra* discussion Section IV.D.2.

226. *See supra* discussion Section II.C.

227. State v. Loomis, 881 N.W.2d 749, 769–70 (Wis. 2016).

VI. CONCLUSION

Algorithms can be powerful tools to improve upon human decision-making. Algorithms, however, are not "Magic 8 Balls" that inevitably achieve that improvement. All algorithms have error rates that can produce potentially significant numbers of false positives and negatives, which varies depending on the use to which an algorithm is put. An algorithm that is "highly predictive of [healthcare] cost[s]"[228] may be significantly more accurate at making those predictions based on a population's gender and age range than based on its health conditions.[229] And the use of an algorithm for a secondary purpose (not the original purpose for which the algorithm was designed for), even one that is "highly correlated [with costs]"[230]—such as predicting a patient's health conditions—can give rise to adverse impacts, such as "exhibit[ing] significant racial bias."[231] The secondary use can also give rise to perverse outcomes, where an algorithm intended to identify patients in need of proactive medical attention de-prioritizes those very patients.[232] We are past the point where organizations can rely solely on their unaided intuition to ensure that the delegation of decision-making to an algorithm will avoid foreseeable adverse consequences.

The good news is that healthcare organizations are well positioned to address these challenges. Hospitals and health systems frequently have medical officers who oversee clinical operations and are versed in a wide range of issues surrounding healthcare delivery. They also possess institutional research and data science expertise that can be utilized to validate how an algorithm functions before and after it is deployed in clinical operations. They can marshal these resources to conduct thorough risk assessments of a decision-making algorithm's implementation to anticipate the expected number of false positives and negatives and the likelihood that those errors adversely impact any protected classes or result in perverse outcomes. That assessment, in turn, can inform the algorithm's implementation so that those foreseeable adverse consequences can be avoided.

---

228. Snowbeck, *supra* note 147.

229. *See* discussion *supra* Section IV.D.4.

230. *See* Dissecting Bias, *supra* note 1.

231. *Id.* at 1.

232. *See* discussion *supra* Section IV.B.

As discussed, the hospital described in *Dissecting Bias* may well have applied a similar approach when it implemented ImpactPro.[233] The next step would be for such approaches to be discussed openly so that they can evolve into best practices or a formalized risk management process aimed at addressing the inherent risks of algorithmic decision-making. In this regard, healthcare organizations could offer a model for the many other industries confronting the same inherent risks of algorithmic decision-making.

233. *See supra* discussion Section II.B.1.