Algorithmic Bias Playbook

Ziad Obermeyer
Rebecca Nissan
Michael Stern
Stephanie Eaneff
Emily Joy Bembeneck
Sendhil Mullainathan

June, 2021

CHICAGO BOOTH | The Center for Applied Artificial Intelligence

ALGORITHMIC BIAS PLAYBOOK

Is your organization using biased algorithms? How would you know? What would you do if so?
This playbook describes 4 steps your organization can take to answer these questions. It distills insights from our years of applied work helping others diagnose and mitigate bias in live algorithms.

Algorithmic bias is everywhere. Our work with dozens of organizations—healthcare providers, insurers, technology companies, and regulators—has taught us that biased algorithms are deployed throughout the healthcare system, influencing clinical care, operational workflows, and policy.

This playbook will teach you how to define, measure, and mitigate racial bias in live algorithms. By working through concrete examples—cautionary tales—you'll learn *what bias looks like*. You'll also see reasons for optimism—success stories—that demonstrate how *bias can be mitigated*, transforming flawed algorithms into tools that fight injustice.

Who should read this? We wrote this playbook with three kinds of people in mind.

- *C-suite leaders (CTOs, CMOs, CMIOs, etc.)*: Algorithms may be operating at scale in your organization—but what are they doing? And who is responsible? This playbook will help you think strategically about how algorithms can go wrong, and what your technical teams can do about it. It also lays out oversight structures you can put in place to prevent bias.
- *Technical teams working in health care*: We've found that the difference between biased and unbiased algorithms is often a matter of subtle technical choices. If you build algorithms, this playbook will help you make those choices better. If you purchase or apply them, it will make you a more 'educated consumer' who can identify problems before they scale.
- *Policymakers and regulators* need to clearly define what algorithmic bias looks like. This playbook's practical approach to bias, which parallels discrimination law, can be used to craft prospective guidance for industry, or to guide retrospective civil investigations.[1]

How do we define 'bias'? There are many definitions of algorithmic bias.[2] We use a practical one, grounded in the real-world use cases of algorithms we've encountered. In health care, we are often faced with a limited supply of resources: tests, treatments, or other forms of care or extra help. Algorithms are used to help decision-makers identify who needs these resources. More generally, in many important social sectors, algorithms guide decisions about who gets what. In these situations, we believe that if an algorithm scores two people the same, those two people should have the same basic needs—no matter the color of their skin, or other sensitive attributes. (This is related to 'calibration' in the literature.) We consider algorithms that fail this test to be biased.

[1] Robert P. Bartlett et al. "Algorithmic Discrimination and Input Accountability under the Civil Rights Acts." Available at SSRN 3674665 (2020).
[2] There is a wealth of literature on this topic. Good starting points are: Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness in machine learning*. fairmlbook.org, 2019; Irene Y. Chen et al. "Ethical Machine Learning in Healthcare." *Annual Review of Biomedical Data Science* 4 (2020); Alvin Rajkomar et al. "Ensuring fairness in machine learning to advance health equity." *Annals of internal medicine* 169, no. 12 (2018): 866-872; Harini Suresh and John V. Guttag. "A framework for understanding unintended consequences of machine learning." *arXiv*:1901.10002 (2019).

How do we measure bias? The first step is always to define the problem the algorithm is trying to solve: What information would an ideal algorithm provide? If the algorithm aims to predict or measure a health need: what is that need, exactly? If the algorithm feeds into a specific decision (who gets a test, a treatment, or some form of extra help): what would be most useful to guide the decision-maker? How to answer these questions in actual data is an important topic that we'll revisit later (health datasets contain a variety of rich measures we can draw on). But before we ever touch the data, we need to articulate the *ideal target* for the algorithm. That ideal target embodies our value system: what do we want the algorithm to learn? It will be the yardstick we use to hold the algorithm accountable, both in general and especially for underserved groups.

What causes bias? Research has identified many causes of algorithmic bias.[3] For our purposes we'll distinguish between two broad categories. The first is when algorithms are aimed at the right target, but fail to hit it for underserved groups. This is often because they were trained or evaluated in non-diverse populations. For example, pulse oximeter devices measure light absorption through the skin, and pass the resulting data through an algorithm to approximate a patient's true blood oxygen levels. Though the algorithm was focused on the right problem, Sjoding et al. recently showed that it performed poorly in Black patients, likely because it was trained on primarily White patients.[4] (This is related to 'representation' and 'evaluation' biases in the literature.) This violates the principle that patients with the same score should have the same true need or outcome, irrespective of race.

Throughout our work on algorithmic bias, though, we've found that a second category is far more common: algorithms are *aimed at the wrong target* to begin with. The result is an insidious 'label choice bias,' arising from a mismatch between the ideal target the algorithm *should be predicting*, and a biased proxy variable the algorithm *is actually predicting*. (This is related to 'measurement bias' in the literature.) For example, we studied a family of algorithms that aim to identify patients with complex health needs, in order to get them extra care.[5] Here, the ideal target is patients' future health needs. But what does that mean, concretely? Algorithms are extremely literal—they predict a specific variable, in a specific dataset—and there is no one variable called 'future health needs.' So instead, many algorithms were trained to predict a proxy variable that *is* present in our datasets: future healthcare costs. Costs seem like a reasonable proxy for health needs. After all, sick people generate health costs. But because of discrimination and barriers to access, Black patients who need health care are less likely to get it, resulting in lower costs—and making costs a racially biased proxy for needs. This mismatch resulted in enormous racial bias in cost-prediction algorithms, affecting important medical decisions for tens of million people every year in the US. The cause of bias was different from pulse oximeters, but the end result was the same: two patients with the same score had very different needs, depending on race.

---

[3]Barocas, Hardt, and Narayanan "Fairness"; Chen et al., "Ethical Machine Learning"; Rajkomar et al., "Ensuring Fairness"; Suresh and Guttag, "A Framework for Understanding."

[4] Michael W. Sjoding, et al. "Racial bias in pulse oximetry measurement." New England Journal of Medicine 383, no. 25 (2020): 2477-2478. This is similar to Joy Buolamwini's work, which has shown that facial recognition algorithms trained in non-diverse samples fail to generalize, affecting performance. See Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Proceedings of Machine Learning Research*: 81:1-15, 2018.

[5] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366, no. 6464 (2019): 447-453.

Are there automated checks I can run to detect bias? When algorithms *are predicting the ideal target*, like in the pulse oximeter example above, basic checks can suggest or confirm bias: under-representation of underserved groups in the training data, or poor accuracy in a way that fits the definition of bias above. These checks can be informative—but only if the algorithm's actual target matches the ideal target. If not, you should not be reassured by good performance.

Unfortunately, no basic checks will tell you when algorithms *are not predicting the ideal target*. This is why label choice bias often goes undetected. In our example above, where cost was being used as a proxy for health needs, basic checks would have shown that the algorithm was working well, for the narrow task it was asked to do: predicting cost, which it did accurately for Black and White patients alike. That was the problem — it predicted a biased target very well. Had we been falsely reassured by this fact, we would have missed large-scale label choice bias. The only way to reveal label choice bias is for a human to articulate the ideal target, and hold the algorithm accountable for that.

Can biased algorithms be fixed? Defining an algorithm's ideal target is at the core of our definition of bias. It can also be a blueprint for improving biased algorithms: once we know what the algorithm should be doing, we know how to retrain the algorithm to do better. If the cause is non-representative training data or failure to generalize, the algorithm can be improved with better data. If the cause is label choice bias, the algorithm can be retrained, to predict a variable closer to its ideal target. In our work, we have learned that the re-trained algorithms are far more fair: they get resources to those who need them, not those who are already well-represented in data. But they also just work better, for everyone: they better match the purpose they were actually designed for.

Is this playbook specific to health care? Health care is a good 'model system' to study algorithmic bias: algorithms operate at a massive scale and can be studied on the servers of a diverse set of organizations. For this reason, our examples come from health care—but the lessons we've learned are very general. We have applied them in follow-on work in financial technology, criminal justice, and a range of other fields.[6] We've found that label choice bias in particular is common in these settings too: for example, finance datasets don't have a variable called 'creditworthiness,' but they do have 'income'; criminal justice datasets don't have a variable called 'criminality,' but they do have 'arrests' and 'convictions.' All of these proxy variables are distorted, biased versions of the ideal target, and similar problems—and solutions—apply.

How do I get started? Our framework is simple and practical, and involves four steps:

- **STEP 1: INVENTORY:** List all the algorithms being used or developed in your organization.
- **STEP 2: SCREEN:** Screen each algorithm for bias, relative to its ideal target.
- **STEP 3: RETRAIN:** Improve or suspend the use of biased algorithms.
- **STEP 4: PREVENT:** Set up structures to prevent future bias.

[6] For criminal justice applications, see also Kristian Lum and William Isaac. "To predict and serve?." *Significance* 13, no. 5 (2016): 14-19.

ALGORITHMIC BIAS CHEAT SHEET

How to use this checklist: This outline of our framework is intended to help you navigate this document and guide your approach to the algorithms in your own institution. You can find detailed instructions, research sources, and case studies by following the links to the appropriate sections.

Step 1: Inventory Algorithms	
	Step 1A: Talk to relevant stakeholders about how and when algorithms are used: Create a list of algorithms within your organization; consider broad definitions of algorithms and ask open ended questions.
	Step 1B: Designate a 'steward' to maintain and update the inventory: Choose a person to be responsible for keeping the inventory current, in consultation with a diverse group.
Step 2: Screen for Bias	
	Step 2A: Articulate the ideal target (what the algorithm should be predicting) vs. the actual target (what it is actually predicting): Consider whether there is a mismatch that can cause bias.
	Step 2B: Analyze and interrogate bias: Choose comparison groups (e.g. race), and perform some basic checks of how well the algorithm predicts its *actual* target. Then, investigate how label choice might create bias in how well the algorithm predicts its *ideal* target.
Step 3: Retrain Biased Algorithms (or Throw Them Out)	
	Step 3A: Try retraining the model on a label closer to the ideal target: Assess possible mitigations to label choice bias by comparing results between different labels.
	Step 3B: Consider alternative options (if necessary): If you are unable to improve or retrain the algorithm, consider other possible solutions. If data is the problem — a non-representative dataset, or no variables that match the ideal target — consider collecting new data.
	Step 3C: Consider suspending or discontinuing use of the algorithm (if necessary): If you are unable to improve the algorithm and/or its inputs, pause the use of the algorithm until you find a solution — or discontinue use altogether.
Step 4: Set Up Structures to Prevent Future Bias	
	Step 4A: Implement best practices for organizations working with algorithms: Under the aegis of the steward and a diverse team, conduct recurring audits and ensure rigorous documentation of current and future models.

ALGORITHMIC BIAS AUDIT PROCESS GUIDE

STEP 1: Inventory Algorithms

The first step is simply making a list of all the algorithms your organization currently has in use or in development. This 'algorithm inventory' will serve as the crucial backbone for subsequent audit steps. Populating the inventory will help unearth algorithms that are most likely to contain bias and thus the highest priority for audits (the prioritization process is further detailed in Step 2A).

But sometimes, it can be hard to know where to start - there is often no central place to find algorithms, the algorithms themselves could have different owners and business purposes, and you may not have access to a log of previously deployed algorithms. So where to begin?

Step 1A: Talk to relevant stakeholders about how and when algorithms are used

Before talking to anyone, it's important to consider the question: what is an algorithm? We define an algorithm broadly, as any quantitative information that is presented to a decision-maker with the goal of informing a decision. Some commonly used algorithms are simple rule-based systems, like the APACHE II score to estimate mortality risk. Others are sophisticated models that summarize large volumes of data, like a patient's entire history of insurance claims or electronic health records, to predict future cost trajectories (we revisit this in detail below). Still others are machine vision models that 'read' imaging data like x-rays or ECGs. These algorithms are increasingly used in clinical settings, to identify patients at high risk of diagnoses like diabetes or delirium. But as we've learned from our work, they are already very widespread for operational or policy decisions that involve allocating scarce resources: flagging patients who are likely to no-show for a scheduled appointment, or deciding who gets access to 'extra help' programs.

With these considerations in mind, talking to a diverse group of decision-makers across business units in your organization is critical. Make a list of key decision-makers who use data—clinicians, researchers, managers, administrators, engineers, IT specialists, and patients—and talk to as many as you can, using structured interviews or more informal conversations. Keep in mind that some stakeholders may not be fully attuned to their use of risk scores or algorithms, or they might not even categorize a tool they use as an algorithm. Keep initial conversations broad and ask as open-ended questions as possible. Choose an interviewer with excellent communication skills who has experience working in similar contexts as the individuals they question. Understanding the kinds of decisions being made and the sorts of tools used to aid those decisions is key to finding what algorithms are being used.

Tip: Search central databases or health records for keywords that relate to algorithms

For example, algorithm outputs such as clinical risk scores may be stored in Electronic Health Records (EHR) alongside other clinical and laboratory data. In that case, string searches for variables containing "score", "scale", "screen", "assess", "index", "tool", "risk", "predict", "model", "algorithm" may help highlight existing repositories of algorithm scores. In other contexts, algorithm outputs may be archived in existing internal databases or databases maintained by partners. Once scores from a new algorithm are identified in this search, proactive outreach to stakeholders can provide additional context for how the algorithm is used to make decisions.

OUTPUT OF STEP 1A: An inventory listing all algorithms your organization is currently using or developing (see example).

Step 1B: Designate a 'steward' to maintain and update the inventory

Someone needs to take responsibility for algorithmic oversight.[7] Developing and maintaining algorithm inventories will require active upkeep and should be overseen by a centralized person. Since algorithms impact entire organizations, the steward should have oversight on broad strategic decisions (i.e., somebody in the C-suite). While this individual will shoulder responsibility for this effort, they should not work alone, but rather in close collaboration with a diverse committee of internal and external stakeholders.

Engaging Communities to Support Bias Mitigation Efforts

Community stakeholders offer valuable input on frameworks for bias mitigation. Involving them in the creation of these structures facilitates transparency and builds trust. For example, we worked with a health plan whose existing Healthcare Ethics Program organized a committee of diverse stakeholders — providers, employers, policy makers, and, crucially, patients themselves. The committee generated a framework that people could reference when using or procuring algorithms and iterated on that framework to ensure it was functional and practical.

We provide further detail on this group's long term responsibilities in Step 4, but forming such a group should be a thoughtful exercise undertaken in consultation with affected communities.The team should include people from different racial groups, genders, etc. and people with diverse areas of expertise such as clinicians, data scientists, business analysts, bioethicists, social science researchers and others. We also encourage a particular focus on representation of members of groups you anticipate focusing on in analyses. For more in-depth information on how to assemble diverse teams and create a culture that promotes responsible AI, we recommend the resources provided by the Center for Equity, Gender, and

[7] Stephanie Eaneff, Ziad Obermeyer, and Atul J. Butte. "The case for algorithmic stewardship for artificial intelligence and machine learning technologies." Jama 324, no. 14 (2020): 1397-1398.

Leadership at the University of California, Berkeley. Their guide on mitigating bias in artificial intelligence is a perfect complement to our work.

In addition to overseeing the algorithm inventory in its current form, the steward and their team may also want to consider adding past algorithms no longer in deployment, and/or ideas for future algorithms (and labeling them accordingly) to make the inventory even more comprehensive.

OUTPUT OF STEP 1B: A designated steward and an oversight structure for algorithms and algorithmic bias.

STEP 2: Screen for Bias

An old programming adage defines debugging as: figuring out what you told the computer to do, as opposed to what you thought you told it to do. This is how we think of Step 2: debugging algorithms. And as any programmer will tell you, debugging requires careful, meticulous work. To build intuition, we'll work through one row of the inventory—an example of an algorithm your organization might be using.

Step 2A: Articulate the algorithm's ideal target vs. its actual target

A good place to start is where we started: our first study on algorithmic bias in health, which we have since replicated in other settings, with other partners.[8]

Imagine you work in an accountable care organization (ACO: a health system that takes responsibility for both the medical care and finances of their patients). One of the algorithms that came up in the inventory is used by the population health team. The team lead describes it as a "risk engine" that helps them better understand their patient population. Your first task in this step is to articulate the *actual target*, the variable that the algorithm actually predicts. You ask them directly which specific variable the algorithm is predicting. They are a bit confused by the question, and answer by saying that it predicts risk, and helps them identify patient groups that need attention. You are confused yourself, until you carefully review the algorithm developer's promotional materials. You learn that, concretely, the model predicts a patient's healthcare costs over the next year. This is the *actual target*: total one-year medical expenditures.

Checking how well the algorithm predicts its actual target is important. Below we'll cover a few basic checks you can do, that can indicate poor performance in an underserved group. If you see this, you should suspect bias. But a key learning from our work is that *accurate prediction of the actual target doesn't guarantee fairness*: the actual target can itself encode biases. Indeed, that is the most common mechanism

[8] Ziad Obermeyer et al., "Dissecting Racial Bias."

of algorithmic bias we've encountered in our work over the past few years. That's why it's so important to step back from the actual target, and articulate the *ideal target*: what the algorithm should be predicting. These two can be very different.

To articulate that ideal target, you go back to your contact and ask what decision, exactly, the algorithm helps them make. After some back and forth, you understand that the algorithm is used to identify patients whose health is likely to get worse, so that they can be enrolled in an 'extra help' program (high-risk care management). The top few percent of patients are fast-tracked into the program, and the bottom percentiles are screened out. You now know enough to articulate the *ideal target*: health care needs, for which extra help and attention can make a difference.

When the actual target fails to match the ideal target, you should worry about bias. In this case, health care costs are being used as a proxy measure of health care needs. Even though cost and health needs are correlated, two patients with the same level of need might have different costs if one receives less care. In fact, less money is spent on Black patients and other populations of color relative to equally sick White patients with the same level of need. This is in part due to barriers that Black populations disproportionately face related to access to care, such as underinsurance or lack of reliable transportation to a nearby hospital. There is also evidence shown by Schulman et al. that doctors treat Black patients differently, recommending less care than they would for White patients.[9]

The way we define an algorithm's target is a manifestation of our value system, as researchers are increasingly pointing out.[10] That is why the seemingly reasonable assumption that cost is a proxy for need is so dangerous: it values people who *get* health care more than people who *need* health care. At any given risk score, general indicators of health needs (e.g. the number of active chronic conditions) for Black patients should be the same as those for White patients; put another way, Black patients in the 90th percentile of algorithmic risk should need care at the same level as White patients in the 90th percentile. Unfortunately, this is not the case for the applications we examine — instead, we see that for a given high risk percentile, Black patients are reliably and systematically sicker than White patients.

Because the algorithm was not predicting its ideal target — health needs — Black patients were deprioritized. They received less care despite a similar level of health. That conclusion is even more concerning because these examples represent only two of many algorithms that payers use to predict costs.[11]

It's natural to blame algorithms for the kinds of biases we describe here: their predictions, on health costs or any other proxy metrics with embedded biases, reinforce and perpetuate systemic racism. But is this really the algorithm's fault? We told it to predict cost, and that's what it did — we didn't tell it that what we actually cared about was health. This reminds us of the 'literal genie' joke genre, for example, when someone tells a genie the first of three wishes: "I want to be rich!" The genie responds, "Okay, Rich. What is your second

[9] Kevin A. Schulman et al. "The effect of race and sex on physicians' recommendations for cardiac catheterization." *New England Journal of Medicine* 340, no. 8 (1999): 618-626.
[10] Samir Passi and Solon Barocas. "Problem Formulation and Fairness." In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pp.39-48, 2019; Maximilian Kasy and Rediet Abebe. "Fairness, Equality, and Power in Algorithmic Decision-Making." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 576-586, 2021.
[11] Geof Hileman and Spenser Steele. "Accuracy of claims-based risk scoring models." Society of Actuaries, 2016.

wish?" Algorithms are literal genies - they give us exactly what we ask for, even if we meant something very different. That's why it's so important to ensure that our actual target variable matches our ideal target as closely as possible.

The subtle but pernicious discrepancy between cost and health needs is just one example of label choice bias as a broader phenomenon. The table below details just a few of the many examples we've found throughout our collaborations with large organizations including hospital systems, for- and non-profit insurers, state and federal agencies, software companies, and others.

Example: Screening for Label Choice Bias

Algorithm	Ideal Target	Actual Target	Risk of Bias
Care Management Prioritization: Identifying patients for additional services	Health needs, benefit from high-risk care management programs	Total costs of care	High. Less money is spent on Black patients who have the same level of need
Emergency Severity Index (ESI): emergency triage	Medical condition needing immediate attention	Nurse-rated acuity, "resources patient is expected to consume"	High. Resource consumption varies by race and insurance for any given acuity
6-Clicks Mobility Score: Decisions about discharge destination	Inability to care for self and live independently at home without help	Physical measures of mobility and daily activities	High. Similar physical mobility scores have larger impact on those lacking income
"No-show" prediction: Clinic scheduling	Voluntary no-show to appointment	Any no-show to prior appointment	High. No shows relate to access: barriers are unequally distributed
Predicting Disease Onset: Targeting preventative care	New disease onset (e.g., heart failure, kidney failure)	Provider–insurer transaction with ICD code for disease	High. Probability of being coded varies by physician quality, hospital billing, insurance, etc.
Kellgren-Lawrence Grade: Osteoarthritis on knee x-rays	Severity of knee osteoarthritis	Severity of osteoarthritis seen by radiologist on knee x-rays	High. Radiologists miss causes of knee pain affecting underserved groups

Table 1[12]

[12] Sendhil Mullainathan and Ziad Obermeyer. "On the Inequality of Predicting A While Hoping for B." *AER Papers and Proceedings* 111:37-42.

In the course of screening for label choice bias, organizations should fill out a table much like this, using the last column to determine the extent to which the discrepancy between the ideal and actual target is likely to create bias for underserved groups. In our running example, which is presented in the first row, we have known for decades that health spending – conditional on need – is lower for Black patients than for White patients.[13] This means there is high risk for bias in an algorithm that predicts cost when the ideal target is need – and indicates that we should prioritize this algorithm in step 2B.

How Label Choice Bias Relates to Discrimination Law[14]

The Supreme Court's 1977 decision in Dothard v. Rawlinson ruled against a prison system's minimum height and weight requirement for hiring. The prison was using these characteristics as proxies for strength, which was required for the job. But because they used proxies—not actual strength tests—the Court ruled they were discriminating against female applicants.

OUTPUT OF STEP 2A: A 4-column table detailing algorithm name, ideal target, actual target, and hypothesized risk of bias

Step 2B: Analyze and interrogate bias

After getting the lay of the land, you're ready to choose a high priority algorithm for further study.

Choosing Populations of Interest

The first thing you'll want to do is choose comparison groups**.** You might have specific interests in some group comparisons going into the analysis — for example, comparing patients by geography in an area where rural patients face barriers to access, or by language spoken in places where non-English speakers may be underserved. Of course, you should also be aware of protected classes designated by the law such as race, color, religion, national origin, sex, and disability. Additionally, you should consider examining implications for multiple groups that are overlapping or intersectional.[15] Think creatively about the groups within the population you serve that may be subject to bias. Speak to a diverse group of stakeholders to understand their hypotheses of bias and to inform your choices of comparison groups.

[13] José J. Escarce and Frank W. Puffer. "Black-white differences in the use of medical care by the elderly: a contemporary analysis," in *Racial and Ethnic Differences in the Health of Older Americans,* eds. Linda G. Martin and Beth J. Soldo. (Washington, DC: National Academy Press, 1997), pp. 183-209.
[14] Robert P. Bartlett et al. "Algorithmic Discrimination and Input Accountability under the Civil Rights Acts (August 1, 2020)." Available at SSRN: https://ssrn.com/abstract=3674665 or http://dx.doi.org/10.2139/ssrn.3674665.
[15] Buolamwini and Gebru, "Gender shades"; James R. Foulds, et al. "An Intersectional Definition of Fairness." In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918-1921. IEEE, 2020.

What if I don't have the data to identify comparison groups? Articulating the groups you want to identify is a good start, but it's quite common to lack group identifiers. Sometimes this is the result of a choice your organization has made — and that choice should be undone if possible: if your organization is committed to reversing disparities, it needs to know how to measure them. If you lack race information, consider using the Bayesian Improved Surname Geocoding Method (BISG) or other similar validated algorithms that use geographical and surname information to infer race.[16] If you want something simpler, you can also merge in race data by zip code. Additionally, some third parties also offer imputation of race and ethnicity as a service — if you go this route, make sure you trust the source and understand what the algorithm is doing prior to implementation. If you lack information on Socioeconomic Status (SES), consider linking to the American Community Survey (ACS) for block level income or other metrics (while being careful to not let patient data leave your system).

Should I Prevent the Algorithm from Using Race (or Zip Code, etc.)?

Race variables seem like an obvious source of bias — but our work has taught us that the most important source of bias is a different variable: the target the algorithm is predicting**.** If the algorithm is predicting its ideal target, we may want the algorithm to use race and zip code variables: they can help it predict the ideal target better. On the other hand, if the algorithm's actual target is a biased proxy, it will be biased *regardless* of the input variables you include. For example, in our running example on cost and health, the algorithm did not use any race or income variables — but this did not prevent it from accurately predicting cost differences between Black and White patients. The key take away: focus on the target variable the algorithm is predicting — not the variables the algorithm uses to predict it.

Some basic checks: Representation, calibration

With an initial set of comparison groups chosen, you can start the analysis. There is a basic template you can follow here, that we illustrate using our running example on cost prediction.

We'll start with some basic checks on the data and the algorithm to help screen for types of bias that can often be seen in the data you have. These are oriented towards detecting problems related to non-diverse training data, failures to generalize, or simply poor performance in underserved groups.

First, you should compare the training dataset (the population the algorithm learned from) and the population in which the algorithm is being applied. Note that, if you have purchased or repurposed an algorithm from an external group or vendor, you may not be able to access any information on the training dataset, but you can absolutely ask them for this information. Specifically, you should compare two things: a) how underserved groups are defined, and b) the fraction of each underserved group.[17] For (a), be aware that sometimes definitions can be different in the training dataset and your population, creating a set up for

[16] Allen Fremont et al. "When Race/Ethnicity Data Are Lacking." *RAND Health Quarterly*, 2016; 6(1):16. As the authors note, "The BISG method is intended to estimate differences at the group or population level; greater caution should be used in classifying specific individuals' race/ethnicity."
[17] Suresh and Guttag, "A framework for understanding unintended consequences of machine learning."

biases. For (b), note that often, algorithms are trained on non-diverse datasets (because more privileged populations have more data available), but applied in very different settings. If the fraction of Black or female patients, for example, looks very different, you should be on the lookout for poor performance in underserved groups in your population.

Next, we'll do a basic check on whether the algorithm is performing well in underserved groups. This is referred to as 'calibration' in the literature: at a given algorithm score, do patients have the same level of the actual target across underserved groups?

Comparing the Actual Target for Groups of Interest

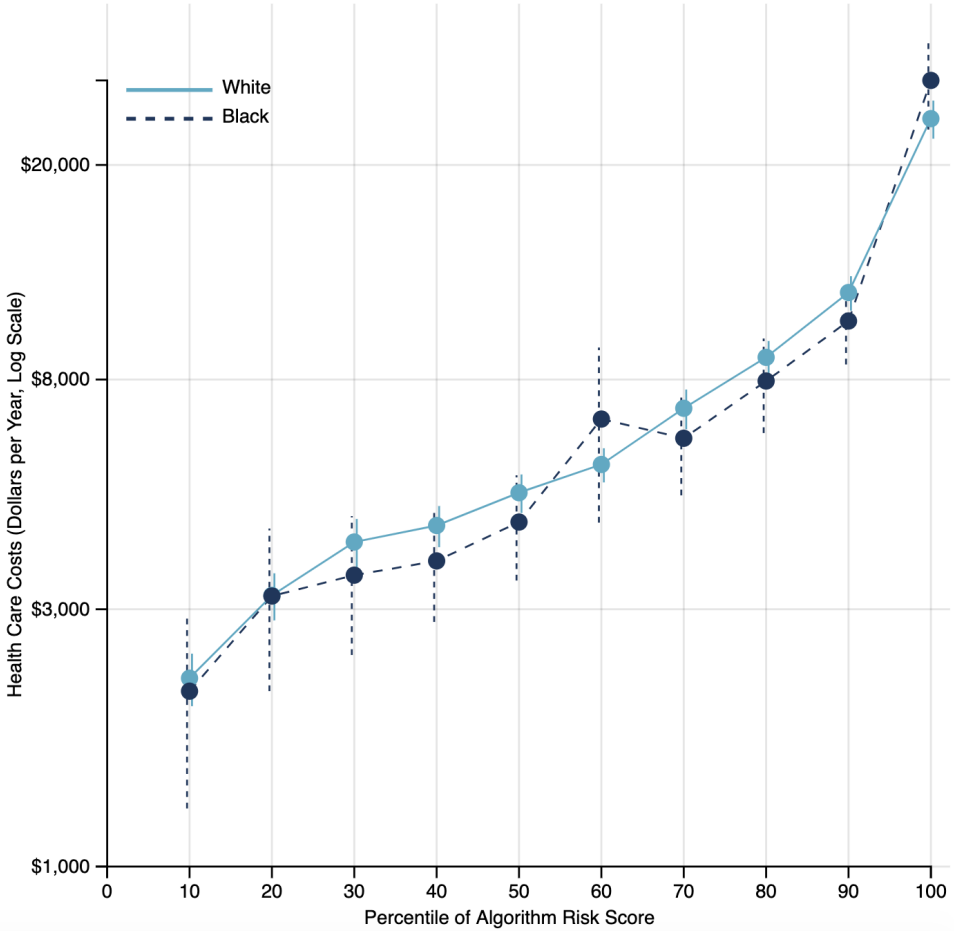

Fig. 1

Notice the performance is good — but don't be reassured. You've ruled out one basic source of bias, namely poor performance for predicting the actual target in an underserved group. But remember, *good performance in an underserved group doesn't guarantee fairness*. You have not ruled out the most common source of bias in the algorithm's we've studied: label choice bias. To do this, you will need to articulate the ideal target — not just take the actual target at face value — and hold the algorithm accountable for predicting that.

Articulating and measuring performance for predicting the ideal target

In the example we walked through in Step 2A, an algorithm is used to prioritize the patients who have the greatest health needs, to get them extra help. In this step, we'll use that example to study how the algorithm (trained to predict the actual target of cost) predicts that ideal target (patient healthcare needs).

Measuring the ideal target. So far we've talked about the ideal target in an abstract way. But now we'll need to get concrete about measuring the ideal target. What do we mean by 'health needs,' exactly? Health is by nature multidimensional and complex — and yet, to quantify bias, we need to measure it precisely, in one or more variables in our dataset. How did we handle this task? We first created an overall measure of health status: the number of active chronic conditions (or "comorbidity score," a metric used extensively in medical research), to provide a comprehensive view of a patient's health.[18] The figure below plots this relationship for the cost-prediction algorithm we have been using as an example:

Difference between Ideal Target and Actual Target by Race

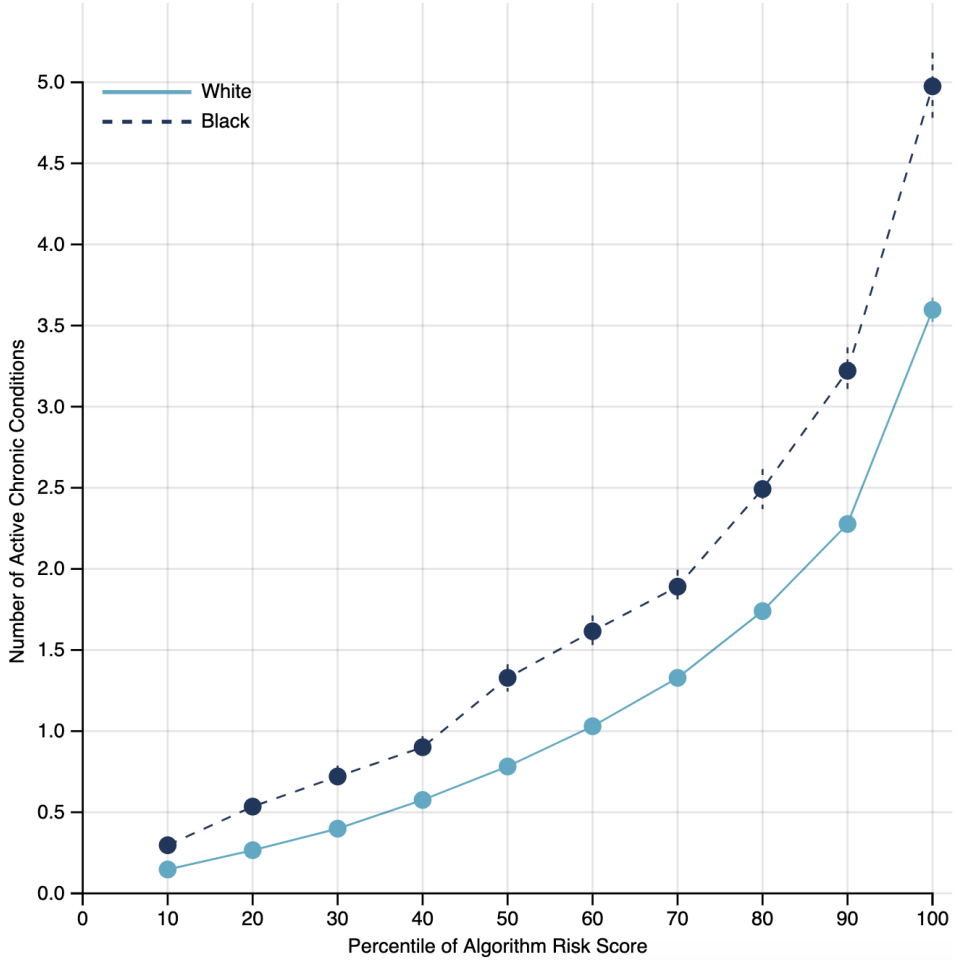

Fig. 2[19]

[18] Vincent de Groot, et al. "How to measure comorbidity: a critical review of available methods." *Journal of Clinical Epidemiology* 56, no. 3 (2003): 221-229.
[19] Confidence intervals for Fig. 2 are present, but narrow, and may be difficult to see when viewed at less than full size.

Here, we see clear evidence of bias, as defined above: for a given risk score, Black patients are sicker than White patients. They had more flare-ups of chronic conditions in the year after the algorithm made its prediction, which guided decisions on who needed extra help. This graph shows that Black patients went on to have worse health, and thus more needs, than a White patient with the same score. In other words, Black patients had to meet a 'higher bar,' in terms of health needs, to get the same help as White patients.

This example offers a generalizable process, and a diagnostic graph, that can be used to check for label choice bias:[20]

1. On the *X-axis*, rank patients by their algorithm score (i.e. predictions on the actual target: Column 3 in the table from Step 2A). You can do this in bins, like percentiles, or deciles.
2. On the *Y-axis*, plot the ideal target (i.e. Column 2 in the table from Step 2A), averaged for all patients in a given bin of algorithm score, separately for each underserved group. If possible, include error bars that capture the 95% confidence interval of each point.
3. If visual inspection is unclear, you can also run a statistical test at some decision threshold applied to the score on the X-axis. For example, if patients in the top 2 percent of scores get enrolled in the extra help program, you can statistically compare the average ideal target in those top 2 percentiles on the Y-axis, between groups.

In practice, you may need to triangulate the ideal target with several different proxy variables, to reflect the complex and multidimensional nature of concepts like 'health' or 'health needs.' To do this, use the same framework: put the candidate ideal target on the Y-axis and the current algorithm score on the X-axis. This lets us assess how the algorithm relates to the candidate ideal target. In the cost prediction example, we compared the algorithm scores to different candidate target variables by swapping out the Y-axis with laboratory values like HbA1c, LDL Cholesterol, Creatinine, Systolic Blood Pressure, and Hematocrit (while leaving the X-axis unchanged). When we did so, we found a similar result: Black patients were still sicker than White patients for most lab values and most percentiles of risk.

An important point here is that all of the metrics we examined are measured with far less bias than health costs. But let's be clear — we are not claiming there is no bias in these measures: to be diagnosed with an active chronic illness, a patient still needs to show up, and barriers to access can intervene. And while blood pressure is measured consistently for different patients, it too is only calculated in patients who show up — not everyone gets measured and we don't know what their blood pressure would have been. This is why it's so important to triangulate the ideal target with as many variables as you can.

Sometimes, it can be hard to generate good candidate targets or to know which one might be the best. If things aren't clear, try to anchor your questions to the decision the algorithm is informing. For the cost prediction example, the decision is: who should I allocate additional care to? Ultimately, which measures you choose will depend on the context in which the algorithm is being used, the data you have available, and input from a broad set of stakeholders. Though it may take several iterations to find the best proxy, the returns on assembling these different measures of the ideal target are enormous — not just for measuring bias in this step, but for retraining less biased algorithms in Step 3 below.

[20] This is very similar to the statistical test proposed by Bartlett et al., in the sense that it quantifies racial differences in the true quantity of interest vs. the proxy.

Measuring the ideal target: another, more nuanced example. A key principle we've learned is that the algorithm's ideal target depends on the decision the algorithm informs. Let's walk through another example that is subtly, but importantly, different from our running example above.

With another one of our partners, a large academic medical center, we worked to study potential bias in triaging patients in their Emergency Department (ED). Triage is a critical part of emergency care: the aim (and the ideal target for any triage algorithm) is to prioritize high-acuity patients for a rapid initial assessment by the medical team. Nearly every ED in the country triages patients using the Emergency Severity Index (ESI), a rule-based algorithm that incorporates (i) a nurse's judgment of acuity, and (ii) a prediction on how many resources a patient was likely to consume in the ED. Our partner worried that resource utilization, a large component of the *actual target*, in particular might be a biased proxy for high-acuity conditions, the *ideal target,* because resource consumption varies by many factors, including race and insurance.

The decision the algorithm informs is which patients get a rapid initial assessment. Notice that this initial assessment is relatively 'cheap': the medical team can always decide that a patient does not need immediate care, and prioritize another patient instead. Because of that, we care much more about making sure patients with critical conditions don't wait than we care whether patients without a critical condition have a (negative) rapid assessment. In other words, we care more about reducing false negatives than we do about reducing false positives. Of course, accurate prediction of both positives and negatives is always an important goal ('calibration'). But the decision context of the algorithm implies that we should be particularly attuned to how often the algorithm misses critical conditions (this is related to metrics of 'recall,' or 'sensitivity'). This is very different from the algorithm above: in that example, prioritizing a patient who doesn't need extra help takes a slot away from another patient who does need it. Extra help is expensive and there are limited slots in the program. So in that setting, we care most about accurate prediction ('calibration'), pure and simple.

How did we measure the ideal target in this case? We convened a group of emergency physicians and experienced nurses to generate a list of the high-acuity conditions they wouldn't want to miss. We then worked with a team of physicians and data scientists to translate that list into a set of diagnoses, laboratory studies, and outcomes that we could measure in the electronic health record data we had. We used that to quantify bias, and show that the existing algorithm did much better for catching critical conditions in White patients than in Black patients. Articulating the ideal target of a triage algorithm is also helping us to lay the groundwork for a better algorithm that corrects some of the problems with ESI and focuses on not missing high-acuity conditions for all patients, irrespective of biases in existing resource use.

OUTPUT OF STEP 2B: A diagnostic chart (or set of key metrics) that illustrates bias in the context of what matters in your specific situation

STEP 3: Retrain Biased Algorithms (or Throw Them Out)

This section is structured as a series of prioritized actions, which we present in the order we typically do them. Some solutions are better than others, so it makes sense to try those first before other options.

Step 3A: Try re-training the model on a label closer to the ideal target

If you made it through Step 2, you've shown bias in an algorithm by comparing its predictions to an ideal target. Now it's time to do something about it. The good news is that much of the hard work is behind you: to fix the biased algorithm, the first thing we try is to retrain it on the same label(s) you used to show bias to begin with — those that match the ideal target.

For example, in the case of the cost prediction algorithm, we found the existing label of cost was biased, by showing that Black patients had far more chronic conditions, higher blood pressure, etc. All of those variables were in our dataset — that's how we were able to show the bias — so mitigating the bias could leverage those same variables. We retrained a new candidate model using active chronic conditions as the label, while leaving the rest of the pipeline intact. This simple change *doubled the fraction of Black patients in the high-priority group: from 14% to 27%*.[21] That said, there are many choices for the alternative label. We could also consider 'avoidable cost' if it was closer to the ideal target for your particular decision and use case. When we did this, it increased the fraction of Black patients identified for the program to 21%. We could train an algorithm to predict a high hemoglobin A1c for diabetes, among those in whom the lab was checked. There are many options, and the best one will depend on your particular circumstances, but the bottom line is that there are often many variables in your datasets that are reasonable proxies for the ideal target.

Before and after making a change, you will want to retrace your analysis to estimate the effect of any given mitigation. Begin by generating a new version of the calibration plot by replacing the old scores with the new model's prediction scores on the x-axis. To complement that number, you can also look at this change of the percent of patients in a given group (e.g., Black patients, non-English speakers, etc.). If your situation is similar to the triage example, looking at your context-specific, prioritized performance metric (e.g. recall) before and after the change may also be useful.

OUTPUT OF STEP 3A: Analysis comparing the level of bias before and after a change OR an assessment that changing the label is infeasible (if the latter, proceed to Step 3B)

[21] Ziad Obermeyer et al. "Dissecting racial bias."

Step 3B: Consider alternative options (if necessary)

Sometimes retraining the model on a less biased label may not be feasible. For example, the data for your alternative label might not be available or algorithms provided by a third-party vendor might not offer this flexibility. In cases like these, there are still other options.[22] It may be possible to purchase access from alternative sources (e.g., commercial datasets from Truven or Optum Labs, or state all-payor claims datasets or national Medicare data). Alternatively, if you are working in a setting like a health system or an insurer, where you can collect more data from patients, that is also an option.

For example, one of our partners was very interested in helping diabetic patients control their blood sugar, by giving them access to dietary and exercise programs. A common approach is to take the entire population of patients, and train an algorithm to predict which will have a high hemoglobin A1c value. But notice there is a problem here: by training the algorithm to find those with high values, *we are implicitly assuming* that those without a hemoglobin A1c value in the system are low. But not everyone who is untested is low: patients who lack access to their doctor might be very high, but they have never been checked. Our partner is solving this by training an algorithm *only* in tested patients: among those with a hemoglobin A1c measured, who will be high? This algorithm can then be used to generate predictions even for patients without a measurement. For the highest-risk patients who have never been tested by their doctors, our partner will send out home testing kits to augment the data and find underserved, undiagnosed diabetics. Preliminarily, the patients the algorithm is finding are disproportionately patients of color and in lower-income neighborhoods, who face many barriers to testing and treatment. By retraining the algorithm to find these patients — not just patients who already have access to testing — and confirming predictions with new data, we are creating a tool that gets help to people who need it.

OUTPUT OF STEP 3B: Analysis comparing the level of bias before and after in relation to your context-specific metrics of interest OR an assessment that alternative changes are insufficient (if the latter, proceed to Step 3C)

Step 3C: Consider suspending or discontinuing use of the algorithm (if necessary)

If none of the solutions described mitigate bias sufficiently (as defined by your context-specific criteria), there is always the option of suspending the use of the algorithm until you find a solution or stopping its use entirely. By this point, you have identified your ideal target, so you can begin the process to procure a new algorithm with an actual target more likely to predict the quantity in which you're truly interested.

[22] Cf. Yoonyoung Park et al. Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. *JAMA Network Open*. 2021;4(4):e213909.

In the next step, we will discuss the organizational structure needed to ensure that newly procured algorithms meet the preventative standards you set.

OUTPUT OF STEP 3C: Suspended or discontinued use of the algorithm and criteria (ideal target) for a new solution

---

STEP 4: Set Up Structures to Prevent Future Bias

So you've read through steps 1-3 above, and you're excited to get started. But in order to operationalize this framework, you need to think big picture: what kind of team do you need to support this work – both immediately and in the long term? As you prepare to audit existing algorithms, it is important to consider how, through this process, you can also create the structures necessary to prevent bias in future algorithms that you create or purchase. Ultimately, bias-prevention practices need to be customized for your organization, but we have included suggestions below based on our experience with a diverse set of partners to help you get started.

Step 4A: Implement best practices for organizations working with algorithms

Organizations working with algorithms should establish protocols for ongoing bias mitigation and set up a permanent team to uphold those protocols.

1. Establish protocols for ongoing bias mitigation. The following systems (at a minimum) should be in place to help your organization consistently and proactively avoid bias:

- ❏ **A pathway for reporting algorithmic bias concerns.** Outline a clear process for *anyone* in the organization to safely report concerns about algorithmic bias to the team without repercussions, and decide on a process for responding to these concerns.
- ❏ **Requirements for documenting algorithms.** Timnit Gebru and others make the point that "in the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information".[23] We should strive to do something similar with algorithms. The team should uphold organization-wide standards for documenting the items below.[24] In order to efficiently track information about your algorithms, consider adding columns with these items directly to your inventory.

[23] Timnit Gebru, et al. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2018).
[24] Cf. Beau Norgeot et al. "Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist." *Nature Medicine* 26(9), pp, 1320-1324.

  - The goal: the algorithm's ideal target, it's actual target, and a bias risk assessment
  - The training process: the data an algorithm employs, its training sample, and how the use-case sample differs from the training sample
  - Performance: the algorithm's performance overall, for underserved groups, and for both the actual and ideal targets
- ❏ **A written plan for regular inventory updates and audits.** Decide on a cadence and/or a set of cues to trigger audits. Since many aspects can change after a model has been deployed, audits should be conducted on a routine basis. Algorithm performance can change when the underlying data change (for example, when a hospital starts using new medical imaging technology) or when algorithms are used in new locations and/or in different contexts (for example, pediatric vs. geriatric or inpatient vs outpatient populations).

2. Assign a permanent team to oversee ongoing bias mitigation efforts. In Step 1B, you designated a steward and diverse group to oversee the algorithmic bias audits efforts you've taken so far. Now it's time to be sure that key responsibilities have been assigned on a permanent basis to sustain the systems and protocols you've developed. The exact structure of the team will vary by organization, but their collective tasks should include those listed below at a minimum. Ask: is someone responsible for each of the tasks listed below?

- ❏ **Address feedback:** Lead the response when a member of your organization identifies a concern related to algorithmic bias.
- ❏ **Check documentation:** Hold members of your organization accountable to documenting all decision-making when creating new models.
- ❏ **Maintain the inventory:** Update the list of algorithms frequently (exactly how often this is necessary will depend on the speed at which your organization develops algorithms).
- ❏ **Instigate audits:** Determine when algorithmic bias audits are necessary, and oversee all audits.

3. Consider working with a third-party to ensure accountability and ongoing guidance. For some organizations, it is helpful to involve a third-party that can oversee or conduct audits. This approach holds organizations accountable and offloads some of the work from your internal team.

4. Stay on top of changes in the field. Keep in mind that this field is developing rapidly, and regulators, quality agencies, and accreditors are increasingly releasing explicit guidelines on the topic (in fact, we are working with several of them), so be sure to look out for future communication on recommendations.

OUTPUT OF STEP 4: Protocols for ongoing bias mitigation and a permanent team responsible for this work

A Note for Policymakers, Accreditors, and Regulators Seeking to Combat Bias

The framework presented above can also be a helpful tool for those who wish to play a role in the guidance and oversight of ethical algorithms. Specifically, we recommend the following:

1. **Prospective Guidance:** Proactively encouraging organizations to follow this framework can help mitigate bias before it occurs. Accreditors can support or certify organizations that prove they have completed an inventory, screened for bias, and mitigated bias where it was found. Certification should also entail adherence to a set of best practices like those listed in Step 4, which represent sustainable organizational structures for bias mitigation.
2. **Retrospective Investigation:** The framework presented in this playbook can also serve as a tool for regulators who want to investigate potential algorithmic bias and hold companies accountable. The inventory is a crucial step: companies need to know what algorithms they are using and how they work. Once such a list exists, regulators can use the table in Step 2A to identify and prioritize algorithms at high risk of bias.

Conclusion

While we have certainly made progress in understanding algorithmic bias, we all still have more to learn. This playbook is intended to be a living document, and we will update it as the field develops. We also welcome questions and feedback. Through ongoing efforts, including a regular conference starting in the fall of 2021, the Center and our partners are committed to continuing to advance both understanding and practice around reducing algorithmic bias.

Acknowledgements

We would like to thank the following individuals for their invaluable feedback on drafts of this playbook:

Mike Berger, Gari Clifford, Ivan Cohen, Chris Hemphill, Ling Hong, Jianying Hu, Doug Jacobs, Annette James, Howard Lakougna, Emma Pierson, Prabhjot Singh, Jann Spiess, Anita Wagner, Thomas Wang, Michael Wilson, and many other colleagues whose ongoing guidance and support has enabled this work. We are especially grateful to our colleagues in public service who remain unnamed, but whose feedback has been invaluable.

Special thanks also to our healthcare industry partners whose collaboration has been instrumental to developing the research insights and best practices you see within this document:

Blue Cross Blue Shield of North Carolina, Brigham & Women's Hospital Department of Emergency Medicine Office of IDEaS, Harvard Pilgrim Health Care, Independence Blue Shield, SymphonyRM, and a host of others who remain unnamed.

In addition, we are grateful to the Center for Applied AI at the University of Chicago Booth School of Business and the University of California Berkeley School of Public Health for continued funding and support.