nature medicine

REVIEW ARTICLE

https://doi.org/10.1038/s41591-021-01614-0

AI in health and medicine

Pranav Rajpurkar[1,4], Emma Chen[2,4], Oishi Banerjee[2,4] and Eric J. Topol[3]✉

Artificial intelligence (AI) is poised to broadly reshape medicine, potentially improving the experiences of both clinicians and patients. We discuss key findings from a 2-year weekly effort to track and share key developments in medical AI. We cover prospective studies and advances in medical image analysis, which have reduced the gap between research and deployment. We also address several promising avenues for novel medical AI research, including non-image data sources, unconventional problem formulations and human–AI collaboration. Finally, we consider serious technical and ethical challenges in issues spanning from data scarcity to racial bias. As these challenges are addressed, AI's potential may be realized, making healthcare more accurate, efficient and accessible for patients worldwide.

In the years ahead, AI is poised to broadly reshape medicine. Just a few years since the first landmark demonstrations of medical AI algorithms that are able to detect disease from medical images at the level of experts[1–4], the landscape of medical AI has matured considerably. Today, the deployment of medical AI systems in routine clinical care presents an important yet largely unfulfilled opportunity, as the medical AI community navigates the complex ethical, technical and human-centered challenges required for safe and effective translation.

In this review, we summarize major advances and highlight overarching trends, providing a concise overview of the state of medical AI. Our review is informed by our efforts over the past 2 years, during which we tracked and shared recent developments in medical AI on a weekly basis (https://doctorpenguin.com). First, we summarize recent progress, highlighting studies that have rigorously demonstrated the utility of medical AI systems. Second, we examine promising avenues for medical AI research in the form of novel data sources and discuss collaboration setups between AI and humans, which are more likely to reflect real medical practice than typical study designs that pit AI against humans. Finally, we discuss major challenges facing the field, including the technological limitations of AI as it stands and ethical concerns about regulating AI systems, holding people accountable when AI error occurs, respecting patient privacy and consent in data collection and safeguarding against the reinforcement of inequities (Fig. 1).

Recent progress in deployment of AI algorithms in medicine

Although AI systems have repeatedly been shown to be successful in a wide variety of retrospective medical studies, relatively few AI tools have been translated into medical practice[5]. Critics point out that AI systems may in practice be less helpful than retrospective data would suggest[6]; systems may be too slow or complicated to be useful in real medical settings[7], or unforeseen complications may arise from the way in which humans and AIs interact[8]. Moreover, retrospective in silico datasets undergo extensive filtering and cleaning, which may make them less representative of real-world medical practice. Randomized controlled trials (RCTs) and prospective studies can bridge this gap between theory and practice, more rigorously demonstrating that AI models can have a quantifiable, positive impact when deployed in real healthcare settings. Recently, RCTs have tested the usefulness of AI systems in healthcare. In addition to accuracy, a variety of other metrics have been used to assess the utility of AI, providing a holistic view of its impact on medical systems[9–13]. For example, an RCT evaluating an AI system for managing insulin doses measured the amount of time patients spent within the target glucose range[14]; a study that evaluated a monitoring system for intraoperative hypotension tracked the average duration of hypotension episodes[15], while a system that flagged cases of intracranial hemorrhage for human review was judged by its reduction of turnaround time[16]. Recent guidelines, such as AI-specific extensions to the SPIRIT and CONSORT guidelines and upcoming guidelines such as STARD-AI, may help standardize medical AI reporting, including clinical trials protocols and results, making it easier for the community to share findings and rigorously investigate the usefulness of medical AI[17,18].

In recent years, some AI tools have moved past testing to deployment, winning administrative support and clearing regulatory hurdles. The Center for Medicare and Medicaid Services, which approves public insurance reimbursement costs, has facilitated the adoption of AI in clinical settings by allowing reimbursement for the use of two specific AI systems for medical image diagnosis[19]. Furthermore, a 2020 study found that the US Food and Drug Administration (FDA) is approving AI, particularly machine learning (ML; a type of AI), products at an accelerating rate[20]. These advances largely take the form of FDA clearances, which require products to meet a lower regulatory bar than full-fledged approvals do, but they are nonetheless clearing a path for AI/ML systems to be used in real clinical settings. It is important to point out that the datasets used for these regulatory clearances are often made up of retrospective, single-institution data that are mostly unpublished and considered proprietary. To build trust in medical AI systems, stronger standards for reporting transparency and validation will be required, including demonstrations of impact on clinical outcomes.

Deep learning for interpretation of medical images. In recent years, deep learning, in which neural networks learn patterns directly from raw data, has achieved remarkable success in image classification. Medical AI research has consequently blossomed in specialties that rely heavily on the interpretation of images, such as radiology, pathology, gastroenterology and ophthalmology.

[1]Department of Biomedical Informatics, Harvard University, Cambridge, MA, USA. [2]Department of Computer Science, Stanford University, Stanford, CA, USA. [3]Scripps Translational Science Institute, San Diego, CA, USA. [4]These authors contributed equally: Pranav Rajpurkar, Emma Chen, Oishi Banerjee. ✉e-mail: etopol@scripps.edu

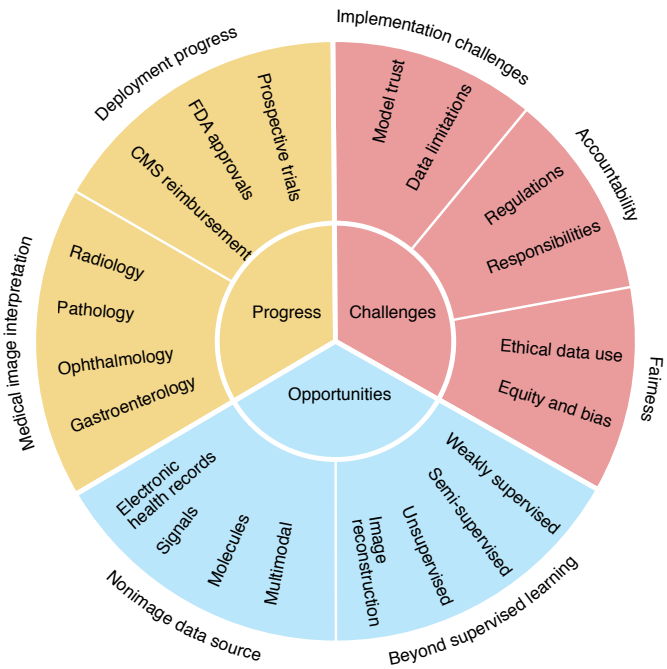

Fig. 1 | Overview of the progress, challenges and opportunities for AI in health. CMS, Centers for Medicare & Medicaid Services.

AI systems have achieved considerable improvements in accuracy for radiology tasks, including mammography interpretation[21,22], cardiac function assessment[23,24] and lung cancer screening[25], tackling not only diagnosis but also risk prediction and treatment[26]. For instance, one AI system was trained to estimate 3-year lung cancer risk from radiologists' computed tomography (CT) readings and other clinical information[27]. These predictions could then be used to schedule follow-up CT scans for patients with cancer, augmenting current screening guidelines. Validation of such systems on multiple clinical sites and an increasing number of prospective evaluations have brought AI closer to being deployed and making a practical impact in the field of radiology.

In the field of pathology, AI has made major strides in diagnosing cancers and providing new disease insights[28–33], largely through the use of whole-slide imaging. Models have been able to efficiently identify areas of interest within slides, potentially speeding up workflows for diagnosis. Beyond this practical impact, deep neural networks have been trained to discern the primary tumor origin and detect structural variants or driver mutations, providing benefits beyond even expert pathologist reviews. Furthermore, AI has been shown to make more accurate survival predictions for a wide range of cancer types compared to conventional grading and histopathological subtyping[31]. Such studies have demonstrated how AI can make pathology interpretations more efficient, accurate and useful.

Deep learning has also made progress in gastroenterology, especially in terms of improving colonoscopy, a key procedure used to detect colorectal cancer. Deep learning has been used to automatically predict whether colonic lesions are malignant, with performance comparable to skilled endoscopists[34]. Additionally, because polyps and other possible signs of disease are frequently missed during the exam[35], AI systems have been developed to assist endoscopists. Such systems have been shown to improve endoscopists' ability to detect irregularities, potentially improving sensitivity and making colonoscopy a more reliable tool for diagnosis[10,11,36].

Deep learning models have been applied widely in the area of ophthalmology, making important advances toward deployment[7,37–41]. Besides quantifying model performance, studies have investigated the human impact of such models on health systems. For example, one study examined how an AI system for eye disease screening affected patient experience and medical workflows, using human observation and interviews[7]. Other studies have looked at the financial impact of AI in the ophthalmology setting, finding that semi-automated[40] or fully automated AI screening[39] might provide cost savings in specific contexts, such as the detection of diabetic retinopathy.

Opportunities for development of AI algorithms

Medical AI studies often follow a familiar pattern, tackling an image classification problem, using supervised learning on labeled data to train an AI system and then evaluating the system by comparing it against human experts. Although such studies have achieved noteworthy advances, we present three other promising avenues of research that break from this mold (Fig. 2). First, we address non-image data sources such as text, chemical and genomic sequences, which can provide rich medical insights. Second, we discuss problem formulations that go beyond supervised learning, obtaining insights from unlabeled or otherwise imperfect data through paradigms such as unsupervised or semi-supervised learning. Finally, we look at AI systems that collaborate with humans instead of competing against them, which is a path toward achieving better performance than either AI or humans alone.

Medical data beyond images. Moving beyond image classification, deep learning models can learn from many kinds of input data, including numbers, text or even combinations of input types. Recent work has drawn on a variety of rich data sources involving molecular information, natural language, medical signals such as electroencephalogram (EEG) data and multimodal data. The following is a summary of applications using these data sources.

AI has enabled recent advances in the area of biochemistry, improving understanding of the structure and behavior of biomolecules[42–45]. The work of Senior et al. on AlphaFold represented a breakthrough in the key task of protein folding, which involves predicting the 3D structure of a protein from its chemical sequence[42]. Improvements in protein structure prediction can provide mechanistic insight into a range of phenomena, such as drug–protein interactions or the effects of mutations. Alley et al. also made strides in the area of protein analysis, creating statistical summaries that capture key properties of proteins and help neural networks learn with less data[43]. By using such summaries rather than raw chemical sequences, models for downstream tasks like predicting molecular function may obtain high performance with much less labeled data.

AI has also made strides in the field of genomics, despite the complexity of modeling 3D genomic interactions. When applied to data on circulating cell-free DNA, AI has enabled noninvasive cancer detection, prognosis and tumor origin identification[46–48]. Deep learning has enhanced CRISPR-based gene editing efforts, helping to predict guide-RNA activity and identify anti-CRISPR protein families[49,50]. Additionally, AI-based analysis of microbial transcriptomic and genomic data has been used to quickly detect antibiotic resistance in pathogens. This advance allows doctors to rapidly select the most effective treatments, potentially reducing mortality and preventing the unnecessary use of broad-spectrum antibiotics[51].

Furthermore, AI is now beginning to accelerate the process of drug discovery. Deep learning models for molecular analysis have been shown to accelerate the discovery of novel drugs by reducing the need for slower, more costly physical experiments. Such models have proven useful for predicting relevant physical properties such as the bioactivity or toxicity of potential drugs. One study used AI to

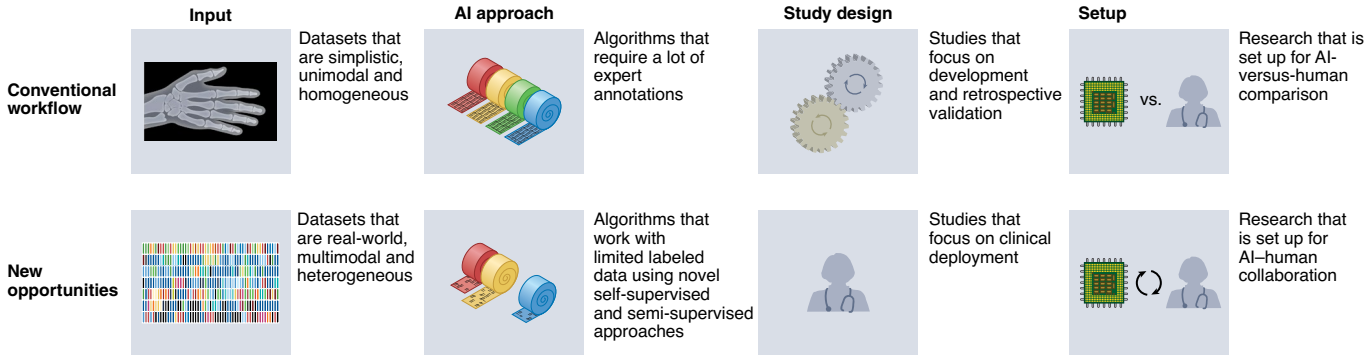

Fig. 2 | Opportunities for the development of AI algorithms.

identify a drug that was subsequently proven to be effective at fighting antibiotic-resistant bacteria in experimental models[52]. Another drug designed by AI was shown to inhibit DDR1 (a receptor implicated in several diseases, including fibrosis) in experimental models; remarkably, it was discovered in only 21 days and experimentally tested in 46 days, dramatically accelerating a process that usually takes several years[53]. Importantly, deep learning models can select effective molecules that differ from existing drugs in clinically meaningful ways, thereby opening novel pathways for treatment and providing new tools in the fight against drug-resistant pathogens.

Recent research has exploited the availability of large medical text datasets for healthcare-related natural language processing tasks, taking advantage of technical advances like transformers and contextual word embeddings (two technologies that help models consider surrounding context when interpreting each part of a text). One study presented BioBERT, a model trained on a large corpus of medical texts that surpassed prior state-of-the-art performance on natural language tasks like answering biomedical questions[54]. Such models have been used to improve performance on tasks such as learning from biomedical literature which drugs are known to interact with each other[55] or automatically labeling radiology reports[56]. Sizable text datasets have also been mined from social media and used to track large-scale mental health trends[57]. Thus, advances in natural language processing have opened up a wealth of new datasets and AI opportunities, although major limitations still exist due to the difficulty of extracting information from long text sequences.

Additionally, ML methods have been used to predict outcomes from medical signal data, such as EEG[58], electrocardiogram[59,60] and audio data[61]. For example, ML applied to EEG signals from clinically unresponsive patients with brain injuries allowed the detection of brain activity, a predictor of eventual recovery[58]. Moreover, AI's ability to directly transform brain waves to speech or text has remarkable potential value for patients with aphasia or locked-in syndrome who have had strokes[62]. Medical signal data can also be passively collected outside a clinical setting in the real world by using wearable sensors such as smartwatches that enable remote health monitoring[59,63].

Some deep learning models integrate multiple sources of medical data for a multimodal approach[64–68]. For instance, one model for diagnosing respiratory disorders took audio recordings of patients' coughs as well as reports of their symptoms as input[65]. Multimodal models have also taken advantage of far more complex inputs, such as electronic health records, which encompass a wide variety of data such as medical diagnoses, vital signs, prescriptions and laboratory results[66,67]. Such models can make predictions based on diverse types of data, much as human clinicians rely on multiple types of information when making decisions in practice. Despite its potential, this area of research seems relatively underdeveloped, in part because of the challenges of gathering multiple types of data consistently across departments or institutions. We nonetheless expect to see the use of multimodal models increase over time.

AI setups beyond supervised learning. In addition to using novel data sources, recent studies have tried unconventional problem formulations. Conventionally, datasets derive inputs and labels from real data, and models like neural networks are used to learn functions mapping from inputs to labels. However, because labeling can be expensive and time-consuming, datasets containing both accurate inputs and labels are often difficult to obtain and are frequently reused across many studies. Other paradigms, including unsupervised learning (specifically self-supervised learning), semi-supervised learning, causal inference and reinforcement learning (Box 1), have been used to tackle problems in which data are unlabeled or otherwise noisy. These advances have pushed the boundary of medical AI, enhancing existing technologies and deepening the understanding of diseases.

Unsupervised learning, which involves learning from data without any labels, has provided actionable insights, allowing models to find novel patterns and categories rather than being limited to existing labels, as in the supervised paradigm[69–73,74]. For example, clustering algorithms, which organize unlabeled data points by grouping similar data points together, have been applied to conditions such as sepsis, breast cancer and endometriosis, identifying clinically meaningful patient subgroups[29,74,75,76]. These categories can reveal novel patterns in disease manifestation that may eventually help to determine diagnosis, prognosis and treatment.

Other formulations rely on extracting information out of noisy or otherwise imperfect data, dramatically reducing the cost of data collection[30,77]. As an example, Campanella et al. trained a weakly supervised model to diagnose several types of cancer from whole-slide images, using only the final diagnoses as labels and skipping the pixel-wise annotation expected in a supervised learning setup. With this approach, they were able to achieve excellent classification results, even with annotation costs lowered[30]. Unconventional problem formulations have also been used to enhance and reconstruct images[78–81]. For instance, when creating a model to enhance spatial detail in low-quality magnetic resonance imaging (MRI) images, Masutani et al. synthetically generated input data; they took high-quality MRI images, randomly added noise and then trained a convolutional neural network (a type of neural network commonly used for image data) to recover the original high-quality MRI images from their simulated 'low-quality' inputs[80]. Such formulations allow researchers to leverage large datasets, despite their imperfections, to train high-performing models.

Setups beyond human versus AI. Although the majority of studies have focused on a head-to-head comparison of AI with humans[82], real-life medical practice is more likely to involve human-in-the-loop setups, where in humans actively collaborate with AI systems and

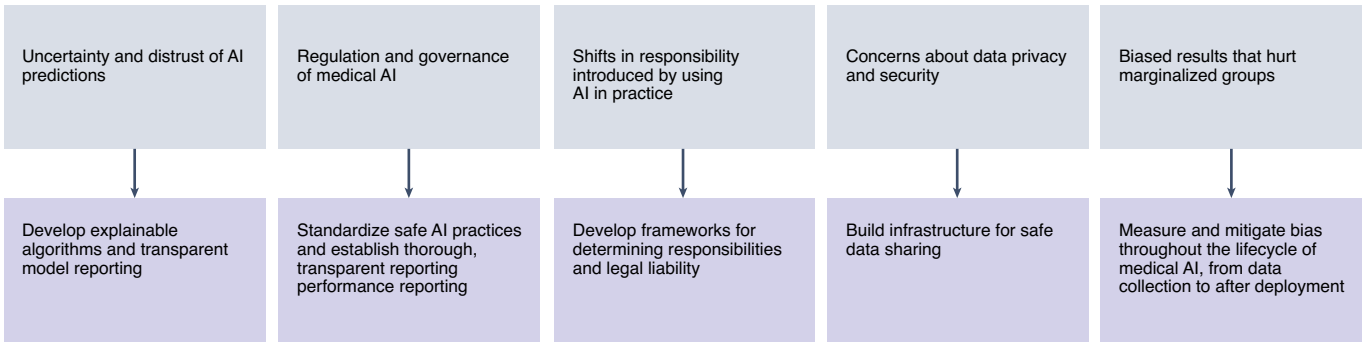

Fig. 3 | Ethical challenges for AI in medicine.

provide oversight[83,84]. Thus, recent studies have begun to explore such collaborative setups between AI and humans. These setups typically feature humans receiving assistance from AI, although occasionally AI and humans work separately and have their predictions averaged or otherwise combined afterward. Multiple studies on a variety of tasks have shown that clinical experts and AI in combination achieve better performance than experts alone[21,85–89]. For example, Sim et al. found that AI-assisted clinical experts surpassed both humans and AI alone when detecting malignant nodules on chest radiographs[85]. The usefulness of human–AI collaboration will likely depend on the specifics of the task and the clinical context.

There are still open questions about exactly how AI assistance affects human performance. For instance, AI assistance has sometimes been shown to improve clinical experts' sensitivity while lowering their specificity[8,86], and some studies, both prospective and retrospective, have found that combined AI–human performance could not surpass the performance of AI alone[90,91]. Furthermore, some clinicians may benefit more from AI assistance than others; studies suggest that less experienced clinicians, such as trainees, benefit more from AI input than their more experienced peers[8,92].

Technical considerations also play a major role in determining the effectiveness of AI assistance. Predictably, the accuracy of AI advice can affect its usefulness, so incorrect predictions have been found to hinder clinician performance even if correct predictions prove helpful[8]. Additionally, AI predictions can be communicated in multiple ways, appearing, for example, as probabilities, text recommendations or images edited to highlight areas of interest. The presentation format of AI assistance has been shown to affect its helpfulness to human users[90,91], so future work on optimizing medical AI assistance may draw on existing research on human–computer interactions.

Challenges for the future of the field

Despite striking advances, the field of medical AI faces major technical challenges, particularly in terms of building user trust in AI systems and composing training datasets. Questions also remain about the regulation of AI in medicine and the ways in which AI may shift and create responsibilities throughout the healthcare system, affecting researchers, physicians and patients alike. Finally, there exist important ethical concerns about data use and equity in medical AI (Fig. 3).

Implementation challenges. *Dataset limitations.* Medical AI data often raise specific, practical challenges. Although it is hoped that AI will reduce medical costs, the devices required to obtain the inputs for AI systems can be prohibitively expensive. Specifically, the equipment needed to capture images of whole slides is costly and is therefore unavailable in many health systems, impeding both data collection for and deployment of AI systems for pathology.

Additional concerns arise from large image sizes, because the amount of memory required by a neural network can increase with both the complexity of the model and the number of pixels in the input. As a result, many medical images, especially whole-slide images, which can easily contain billions of pixels each, are too large to fit into the average neural network. There exist many ways to address this issue. Pictures may be resized at the expense of fine details, or they may be split into multiple small patches, although this will hinder the system's ability to draw connections between different areas of the image. In other cases, humans may identify a smaller region of interest, such as part of a slide image that contains a tumor, and crop the image before feeding it into an AI system, though this intervention adds a manual step into what might otherwise be a fully automated workflow[32,93]. Some studies use large custom models that can accept whole medical images, but running these models can require expensive hardware with more memory. Thus, systems for medical image classification often involve trade-offs to make inputs compatible with neural networks.

Another issue affecting images as well as many other types of medical data is a shortage of the labels required for supervised learning[94]. Labels are often hand-assigned by medical experts, but this approach can prove difficult due to dataset size, time constraints or shortage of expertise. In other cases, labels can be provided by non-expert humans, for example, via crowdsourcing. However, such labels may be less accurate, and crowdsourced labeling projects face complications associated with privacy, as the data must be shared with many labelers. Labels can also be applied by other AI models, as in some weak-supervision setups[95], but these labels again carry the risk of noise. Currently, the difficulty of obtaining quality labels is a major blockade for supervised learning projects, driving interest in platforms that make labeling more efficient and weakly supervised and unsupervised setups that require less labeling effort.

Problems also arise when technological factors lead to bias in datasets. For example, single-source bias occurs when a single system generates an entire dataset, as when all the images in a collection come from a single camera with fixed settings. Models that exhibit single-source bias may underperform on inputs collected from other sources. To improve generalization, models can undergo site-specific training to adapt to the specific quirks of each place they are deployed, and they can also be trained and validated on datasets collected from different sources[94,96]. However, the latter approach must be undertaken with care, especially when the distribution of labels differs dramatically across datasets. For instance, if a model is trained on datasets from two institutions, one containing only positive cases and one containing only negative cases, then it can achieve high performance through spurious 'shortcuts' without learning about the relevant pathology. An image classification model might thus base its predictions entirely on the differences between the two institutions' cameras; such a model would likely learn nothing about the underlying disease and fail to generalize

Box 1 | AI setups beyond supervised learning

Self-supervised learning	Learning from unlabeled data by leveraging information extracted from the data itself
Semi-supervised learning	Learning from a small amount of labeled data combined with a large amount of unlabeled data
Causal inference	Finding the effect of a component or treatment on a system using data
Reinforcement learning	Learning in an interactive environment using feedback from actions and past experiences

elsewhere. We therefore encourage researchers to be wary of technological bias, even when using data from diverse sources[97].

Building model trust. A variety of qualities are desired for an AI system to garner user trust. For example, it is useful for AI systems to be reliable, convenient to use and easy to integrate into clinical workflows[98]. AI systems can be packaged with easy-to-read instructions, explaining how and when they should be used; it may be helpful for such user manuals to be standardized across systems[99].

Explainability is another key aspect of earning trust, as it is easier to buy into an AI system's predictions when the system can explain how it reached its conclusions. Because many AI systems currently function as uninterpretable 'black boxes', explaining their predictions poses a serious technical challenge. Some methods for explaining AI predictions exist, such as saliency methods that highlight regions of an image that most contribute to a prediction of a disease by a model. However, these methods may not be reliable[100], and further research is necessary to interpret AI decision-making processes, quantify their reliability and convey those interpretations clearly to human audiences[101]. In addition to building trust among users, enhanced explainability will allow developers to check models more thoroughly for errors and verify to what degree AI decision-making mirrors expert human approaches[102]. Moreover, when medical AI models achieve novel insights that go beyond current human knowledge, improved explainability may help researchers grasp those new insights and thus better understand the biological mechanisms behind disease.

Perhaps the most obvious component of trustworthiness is accuracy, because users are unlikely to trust a model that has not been rigorously shown to give correct predictions. Additionally, trustworthy AI studies should be reproducible, so that repeatedly training a model with a given dataset and protocol produces consistent results. Studies should also be replicable, so that models perform consistently even when trained with different samples of data. Unfortunately, proving the reproducibility and replicability of AI studies raises unique challenges. Datasets, code and trained models are often not released publicly, making it difficult for the wider AI community to independently verify and build on previous results[103,104].

Accountability. *Regulatory challenges.* Recent work highlights regulatory issues regarding the deployment of AI models for healthcare. Beyond accuracy, regulators can look at a variety of criteria to evaluate models. For example, they may require validation studies showing that AI systems are robust and generalizable across clinical settings and patient populations and ensure that systems protect patient privacy. Additionally, because the usefulness of AI systems can depend heavily on how humans provide input and interpret output, regulators may require testing of human factors and adequate training for the human users of medical AI systems[105].

Specific regulatory challenges arise from continual learning, where models learn from new data over time and adjust to shifts in patient populations, as this may come at the risk of overwriting previously learned patterns or otherwise causing new mistakes[106]. Traditionally, regulators of AI systems approve only one locked set of parameters, yet this approach does not account for the necessity to update models, as data evolve due to changes in patient populations, data collection tools and care management. Regulators must therefore develop novel certification processes to handle such systems. Importantly, the FDA has recently proposed a framework for adaptive AI systems in which they would approve not only an initial model but also a process for updating it over time[107].

Shifts in responsibility. Although AI systems have the potential to empower humans in medical decision-making, they also run the risk of limiting personal autonomy and creating new obligations. As AI systems take on more responsibilities in the healthcare setting, a concern facing the system is that clinicians may become overly reliant on AI, perhaps seeing a gradual decline in their own skills or personal connections with patients. In turn, medical AI developers may gain outsized influence on healthcare and should thus be obliged to create safe, useful AI systems and responsibly influence public views on health. As medical decision-making becomes more reliant on potentially unexplained AI judgments, individual patients might lose some understanding of, or control over, their own care. Patients might at the same time gain new responsibilities as AI makes healthcare more pervasive in daily life. As an example, if smart devices provide patients with constant advice, then those patients may be expected to follow those recommendations or else be responsible for negative health outcomes[108].

The proliferation of AI also raises concerns around accountability, as it is currently unclear whether developers, regulators, sellers or healthcare providers should be held accountable if a model makes mistakes even after being thoroughly clinically validated. Currently, doctors are held liable when they deviate from the standard of care and patient injury occurs. If doctors are generally skeptical of medical AI, then individual doctors may be adversely influenced to ignore AI recommendations that conflict with standard practice, even if those recommendations may be personalized and beneficial for a specific patient. However, if the standard of care shifts so that doctors routinely use AI tools, then there will be a strong medicolegal incentive for doctors to follow AI recommendations[109].

Fairness. *Ethical data use.* There are concerns that bad actors interested in identity theft and other misconduct might take advantage of medical datasets, which often contain large amounts of sensitive information about real patients. Decentralizing data storage is one way to reduce the potential damage of any individual hack or data leak. The process of federated learning facilitates such decentralization while also making it easier to collaborate across institutions without complicated data-sharing agreements (Fig. 4). When using federated learning, developers send AI models to different institutions that have private datasets; the institutions train the models on their data and send back model updates without ever sharing the data[110]. However, even after models are trained, there remains the risk that AI systems will face privacy attacks, which can sometimes reconstruct original data points used in training just by examining the resulting model. Patient data can be better protected from such attacks if inputs are encrypted before training, but this approach comes at the cost of model interpretability[111].

Looking beyond such bad-faith attacks, there are other questions about how to respect patients' privacy. Sensitive data should typically be collected and used in research with patient consent and, where practical, anonymization and aggregation strategies should be used to obscure personal details. It is necessary to ensure that any institutions working with patient data handle them responsibly, for

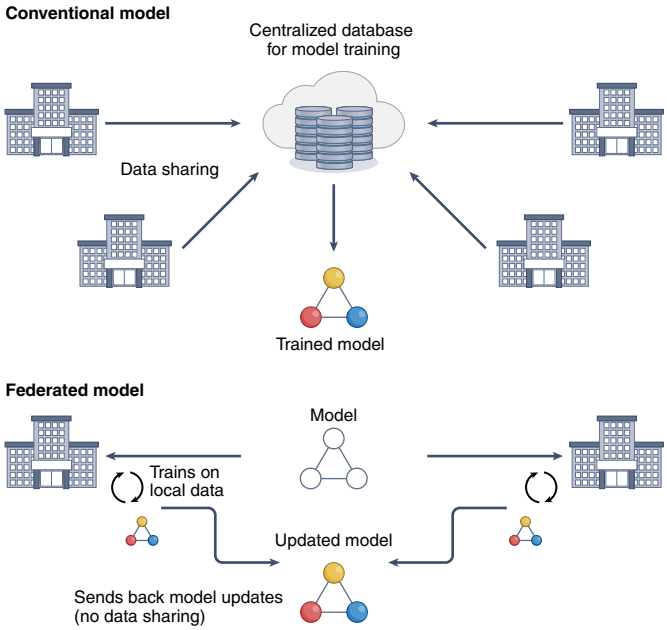

Fig. 4 | Evolving procedures for data sharing. An advantage of federated learning is that it is decentralized, representing a major potential advance in data security.

example, by using appropriate security protocols. At the same time, it is also important that patient data be used for the good of patients. Out of respect for the patients who have agreed to share their personal information, patient data ideally would be used for research that promotes future patient well-being. Unfortunately, these goals can sometimes conflict with each other; implementing security measures like federated learning can require considerable resources and effort, and institutions that cannot make those investments may be unable to access certain datasets, even when their research would benefit the patients in question. Additionally, reusing data across projects may make it more difficult to obtain informed consent, because patients allured by one study may be hesitant to join others. We hope and expect the AI community will continue exploring these trade-offs and find new ways to balance a variety of patient interests[110].

Equity and bias. AI can make healthcare more accessible to underserved groups, but it also risks reinforcing existing inequities, because AI models can perpetuate biases lurking in the data[112]. Medical AI systems can fail to generalize to new kinds of data they were not trained on; thus, training on datasets that underrepresent marginalized groups is well known to result in biased systems that underperform on those groups. Systems that explicitly factor race into their predictions are also at risk of perpetuating prejudice, because racial categories are difficult to define and obscure the diversity within racial groups[113]. Bias can creep in due to other design choices, such as the choice of target label. For example, a risk-assessment algorithm used to guide clinical decision-making for 200 million patients was found to give racially biased predictions, such that white patients assigned a certain predicted risk score tended to be healthier than Black patients with the same score. This bias was due in large part to the original labels used in training. The system was trained to predict future healthcare costs, but because Black patients had historically received less expensive care than white patients due to existing systematic biases, the system reproduced those racial biases in its predictions[114]. Extensive research is necessary to detect and correct bias in medical AI models, because bias can cause widespread harm to marginalized groups if left unchecked. In the future, AI tools may systematically undergo special testing before deployment to verify that neural networks serve the well-being of marginalized populations equitably. Additionally, it may become easier to identify dangerous bias if model explainability improves, because human monitors will be able to double check the reasoning of AI systems and identify problematic elements[115].

Conclusion

The field of medical AI has made considerable progress toward large-scale deployment, especially through prospective studies such as RCTs and through medical image analysis, yet medical AI remains in an early phase of validation and implementation. To date, a limited number of studies have used external validation, prospective evaluation and diverse metrics to explore the full impact of AI in real clinical settings, and the range of assessed use cases has been relatively narrow. Although the field requires more testing and practical solutions, there is also a need for bold imagination. AI has proven capable of extracting insights from unexpected sources and drawing connections that humans would not normally anticipate, so we hope to see even more creative, out-of-the-box approaches to medical AI. There are rich opportunities for novel AI research involving non-image data types and unconventional problem formulations, which open a broader array of possible datasets. Opportunities also exist in AI–human collaboration, an alternative to the AI-versus-human competitions common in research; we would like to see collaborative setups receive more study, as they may provide better results than either AI or humans alone and are more likely to reflect real medical practice. Despite the potential of the field, major technical and ethical questions remain for medical AI. As these pivotal issues are systematically addressed, the potential of AI to markedly improve the future of medicine may be realized.

Received: 23 July 2021; Accepted: 5 November 2021;
Published online: 20 January 2022

References

1. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
2. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
3. Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).
4. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
5. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
6. Kanagasingam, Y. et al. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. *JAMA Netw. Open* **1**, e182665 (2018).
7. Beede, E. et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 1–12 (Association for Computing Machinery, 2020); https://dl.acm.org/doi/abs/10.1145/3313831.3376718
8. Kiani, A. et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit. Med.* **3**, 23 (2020).
9. Lin, H. et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* **9**, 52–59 (2019).
10. Gong, D. et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol. Hepatol.* **5**, 352–361 (2020).
11. Wang, P. et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol. Hepatol.* **5**, 343–351 (2020).
12. Hollon, T. C. et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat. Med.* **26**, 52–58 (2020).

13. Phillips, M. et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw. Open* **2**, e1913436 (2019).
14. Nimri, R. et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nat. Med.* **26**, 1380–1384 (2020).
15. Wijnberge, M. et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs. standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery. *J. Am. Med. Assoc.* **323**, 1052–1060 (2020).
16. Wismüller, A. & Stockmaster, L. A prospective randomized clinical trial for measuring radiology study reporting time on Artificial Intelligence-based detection of intracranial hemorrhage in emergent care head CT. in *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging* vol. 11317, 113170M (International Society for Optics and Photonics, 2020).
17. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Br. Med. J.* **370**, m3164 (2020).
18. Rivera, S. C. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 (2020).
19. Centers for Medicare & Medicaid Services. Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Final Policy Changes and Fiscal Year 2021 Rates; Quality Reporting and Medicare and Medicaid Promoting Interoperability Programs Requirements for Eligible Hospitals and Critical Access Hospitals. *Fed. Regist.* **85**, 58432–59107 (2020).
20. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit. Med.* **3**, 118 (2020).
21. Wu, N. et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* **39**, 1184–1194 (2020).
22. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
23. Ghorbani, A. et al. Deep learning interpretation of echocardiograms. *NPJ Digit. Med.* **3**, 10 (2020).
24. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
25. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
26. Huynh, E. et al. Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* **17**, 771–781 (2020).
27. Huang, P. et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. *Lancet Digit. Health* **1**, e353–e362 (2019).
28. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
29. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
30. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
31. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
32. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
33. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology: new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
34. Zhou, D. et al. Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. *Nat. Commun.* **11**, 2961 (2020).
35. Zhao, S. et al. Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: a systematic review and meta-analysis. *Gastroenterology* **156**, 1661–1674 (2019).
36. Freedman, D. et al. Detecting deficient coverage in colonoscopies. *IEEE Trans. Med. Imaging* **39**, 3451–3462 (2020).
37. Liu, H. et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol.* **137**, 1353–1360 (2019).
38. Milea, D. et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N. Engl. J. Med.* **382**, 1687–1695 (2020).
39. Wolf, R. M., Channa, R., Abramoff, M. D. & Lehmann, H. P. Cost-effectiveness of autonomous point-of-care diabetic retinopathy screening for pediatric patients with diabetes. *JAMA Ophthalmol.* **138**, 1063–1069 (2020).
40. Xie, Y. et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit. Health* **2**, e240–e249 (2020).
41. Arcadu, F. et al. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit. Med.* **2**, 92 (2019).
42. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
43. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
44. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
45. Greener, J.G. et al. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.* **10**, 3977 (2019).
46. Chabon, J. J. et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245–251 (2020).
47. Luo, H. et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Sci. Transl. Med.* **12**, eaax7533 (2020).
48. Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
49. Gussow, A. B. et al. Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat. Commun.* **11**, 3784 (2020).
50. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
51. Bhattacharyya, R. P. et al. Simultaneous detection of genotype and phenotype enables rapid and accurate antibiotic susceptibility determination. *Nat. Med.* **25**, 1858–1864 (2019).
52. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **181**, 475–483 (2020).
53. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
54. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
55. Zhu, Y., Li, L., Lu, H., Zhou, A. & Qin, X. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *J. Biomed. Inform.* **106**, 103451 (2020).
56. Smit, A. et al. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* 1500–1519 (2020).
57. Sarker, A., Gonzalez-Hernandez, G., Ruan, Y. & Perrone, J. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA Netw. Open* **2**, e1914672 (2019).
58. Claassen, J. et al. Detection of brain activation in unresponsive patients with acute brain injury. *N. Engl. J. Med.* **380**, 2497–2505 (2019).
59. Porumb, M., Stranges, S., Pescapè, A. & Pecchia, L. Precision medicine and artificial intelligence: a pilot study on deep learning for hypoglycemic events detection based on ECG. *Sci. Rep.* **10**, 170 (2020).
60. Attia, Z. I. et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* **394**, 861–867 (2019).
61. Chan, J., Raju, S., Nandakumar, R., Bly, R. & Gollakota, S. Detecting middle ear fluid using smartphones. *Sci. Transl. Med.* **11**, eaav1102 (2019).
62. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249–254 (2021).
63. Green, E. M. et al. Machine learning detection of obstructive hypertrophic cardiomyopathy using a wearable biosensor. *NPJ Digit. Med.* **2**, 57 (2019).
64. Thorsen-Meyer, H.-C. et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit. Health* **2**, e179–e191 (2020).
65. Porter, P. et al. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respir. Res.* **20**, 81 (2019).
66. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
67. Kehl, K. L. et al. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncol.* **5**, 1421–1429 (2019).
68. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit. Med.* **3**, 136 (2020).

69. Wang, C. et al. Quantitating the epigenetic transformation contributing to cholesterol homeostasis using Gaussian process. *Nat. Commun.* **10**, 5052 (2019).
70. Li, Y. et al. Inferring multimodal latent topics from electronic health records. *Nat. Commun.* **11**, 2536 (2020).
71. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
72. Li, X. et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 2338 (2020).
73. Amodio, M. et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**, 1139–1145 (2019).
74. Urteaga, I., McKillop, M. & Elhadad, N. Learning endometriosis phenotypes from patient-generated data. *NPJ Digit. Med.* **3**, 88 (2020).
75. Brbić, M. et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods* **17**, 1200–1206 (2020).
76. Seymour, C. W. et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *J. Am. Med. Assoc.* **321**, 2003–2017 (2019).
77. Fries, J. A. et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nat. Commun.* **10**, 3111 (2019).
78. Jin, L. et al. Deep learning enables structured illumination microscopy with low light levels and enhanced speed. *Nat. Commun.* **11**, 1934 (2020).
79. Vishnevskiy, V. et al. Deep variational network for rapid 4D flow MRI reconstruction. *Nat. Mach. Intell.* **2**, 228–235 (2020).
80. Masutani, E. M., Bahrami, N. & Hsiao, A. Deep learning single-frame and multiframe super-resolution for cardiac MRI. *Radiology* **295**, 552–561 (2020).
81. Rana, A. et al. Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis. *JAMA Netw. Open* **3**, e205111 (2020).
82. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
83. Chen, P.-H. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
84. Patel, B. N. et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit. Med.* **2**, 111 (2019).
85. Sim, Y. et al. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* **294**, 199–209 (2020).
86. Park, A. et al. Deep learning–assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Netw. Open* **2**, e195600 (2019).
87. Steiner, D. F. et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am. J. Surg. Pathol.* **42**, 1636–1646 (2018).
88. Jain, A. et al. Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Netw. Open* **4**, e217249 (2021).
89. Seah, J. C. Y. et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit. Health* **3**, e496–e506 (2021).
90. Rajpurkar, P. et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digit. Med.* **3**, 115 (2020).
91. Kim, H.-E. et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Health* **2**, e138–e148 (2020).
92. Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
93. van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
94. Willemink, M. J. et al. Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
95. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. in *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 33, 590–597 (2019).
96. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
97. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
98. Cutillo, C. M. et al. Machine intelligence in healthcare: perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit. Med.* **3**, 47 (2020).
99. Sendak, M. P., Gao, M., Brajer, N. & Balu, S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit. Med.* **3**, 41 (2020).
100. Saporta, A. et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. Preprint at *medRxiv* https://doi.org/10.1101/2021.02.28.21252634 (2021).
101. Ehsan, U. *et al*. The who in explainable AI: how AI background shapes perceptions of AI explanations. Preprint at https://arxiv.org/abs/2107.13509 (2021).
102. Reyes, M. et al. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radio. Artif. Intell.* **2**, e190043 (2020).
103. Liu, C. *et al*. On the replicability and reproducibility of deep learning in software engineering. Preprint at https://arxiv.org/abs/2006.14244 (2020).
104. Beam, A. L., Manrai, A. K. & Ghassemi, M. Challenges to the reproducibility of machine learning models in health care. *J. Am. Med. Assoc.* **323**, 305–306 (2020).
105. Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit. Med.* **3**, 53 (2020).
106. Lee, C. S. & Lee, A. Y. Clinical applications of continual learning machine learning. *Lancet Digit. Health* **2**, e279–e281 (2020).
107. Food and Drug Administration. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback* (FDA, 2019).
108. Morley, J. et al. The debate on the ethics of AI in health care: a reconstruction and critical review. *SSRN* http://dx.doi.org/10.2139/ssrn.3486518 (2019.
109. Price, W. N., Gerke, S. & Cohen, I. G. Potential liability for physicians using artificial intelligence. *J. Am. Med. Assoc.* **322**, 1765–1766 (2019).
110. Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H. & Langlotz, C. P. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology* **295**, 675–682 (2020).
111. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
112. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl Acad. Sci. USA* **117**, 12592–12594 (2020).
113. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight: reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med.* **383**, 874–882 (2020).
114. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
115. Cirillo, D. et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit. Med.* **3**, 81 (2020).

Acknowledgements

We thank A. Tamkin and N. Phillips for their feedback. E.J.T. receives funding support from US National Institutes of Health grant UL1TR002550.

Author contributions

P.R. and E.J.T. conceptualized this Review. E.C., O.B. and P.R. were responsible for the design and synthesis of this Review. All authors contributed to writing and editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Eric J. Topol.

Peer review information *Nature Medicine* thanks Despina Kontos and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Karen O'Leary was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (*1*–*3*). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (*4*), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (*5*), and image searches for professions such as CEO produce fewer images of women (*6*). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (*7*, *8*), and natural language processing algorithms encode language in gendered ways (*9*).

Empirical investigations of algorithmic bias, though, 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 workarounds such as audit studies. Such efforts 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. Our understanding of a mechanism therefore typically relies on theory or exercises with researcher-created algorithms (*10*–*13*). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. 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. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (*14*–*17*). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health systems rely extensively on algorithms to identify patients who will benefit the most (*18*, *19*).

Identifying patients who will derive the greatest benefit from these programs is a challenging causal inference problem that requires estimation of individual treatment effects. To solve this problem, health systems make a key assumption: Those with the greatest care needs will benefit the most from the program. Under this assumption, the targeting problem becomes a pure prediction policy problem (*20*). Developers then build algorithms that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. 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 self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

[1]School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [4]Mongan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. [5]Booth School of Business, University of Chicago, Chicago, IL, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

and ethnic identities. 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 (1 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 (Table 1).

For these patients, we 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. Patients above the 97th percentile are automatically identified for enrollment in the program. Those 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.

Many existing metrics of algorithmic bias may apply to this scenario. Some definitions focus on calibration [i.e., whether the realized value of some variable of interest Y matches the risk score R (*2*, *22*, *23*)]; others on statistical parity of some decision D influenced by the algorithm (*10*); and still others on balance of average predictions, conditional on the realized outcome (*22*). Given this multiplicity and the growing recognition that not all conditions can be simultaneously satisfied (*3*, *10*, *22*), we focus on metrics most relevant to the real-world use of the algorithm, which are related to calibration bias [formally, comparing Blacks B and Whites W, $E[Y|R, W] = E[Y|R, B]$ indicates the absence of bias (here, E is the expectation operator)]. The algorithm's stated goal is to predict complex health needs for the purpose of targeting an intervention that manages those needs. Thus, we compare the algorithmic risk score for patient i in year t ($R_{i,t}$), formed on the basis of claims data $X_{i,(t-1)}$ from the prior year, to data on patients' realized health $H_{i,t}$, assessing how well the algorithmic risk score is calibrated across race for health outcomes $H_{i,t}$. We also ask how well the algorithm is calibrated for costs $C_{i,t}$.

To measure H, we link predictions to a wide range of outcomes in electronic health record data, including all diagnoses (in the form of International Classification of Diseases codes) as well as key quantitative laboratory studies and vital signs capturing the severity of chronic illnesses. To measure C, we link predictions to insurance claims data on utilization, including outpatient and emergency visits, hospitalizations, and health care costs. These data, and the rationale for the specific measures of H used in this study, are described in more detail in the supplementary materials.

Health disparities conditional on risk score

We begin by calculating an overall measure of health status, the number of active chronic conditions [or "comorbidity score," a metric used extensively in medical research (*24*) to provide a comprehensive view of a patient's health (*25*)] by race, conditional on algorithmic risk score. Fig. 1A shows that, at the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites. We can quantify these differences by choosing one point on the x axis that corresponds to a very-high-risk group (e.g., patients at the 97th percentile of risk score, at which patients are auto-identified for program enrollment), where Blacks have 26.3% more chronic illnesses than Whites (4.8 versus 3.8 distinct conditions; $P < 0.001$).

What do these prediction differences mean for patients? Algorithm scores are a key input to decisions about future enrollment in a care coordination program. So as we might expect, with less-healthy Blacks scored at similar risk scores to more-healthy Whites, we find evidence

Table 1. Descriptive statistics on our sample, by race. BP, blood pressure; LDL, low-density lipoprotein.

	White	Black
n (patient-years)	88,080	11,929
n (patients)	43,539	6079
Demographics		
Age	51.3	48.6
Female (%)	62	69
Care management program		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
Care utilization		
Actual cost	$7540	$8442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31
Mean biomarker values		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
Active chronic illnesses (comorbidities)		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Arrythmia	0.09	0.08
Hypothyroid	0.09	0.05
Obesity	0.07	0.18
Pulmonary disease	0.07	0.11
Cancer	0.07	0.06
Depression	0.06	0.08
Anemia	0.05	0.10
Arthritis	0.04	0.04
Renal failure	0.03	0.07
Electrolyte disorder	0.03	0.05
Heart failure	0.03	0.05
Psychosis	0.03	0.05
Valvular disease	0.03	0.02
Stroke	0.02	0.03
Peripheral vascular disease	0.02	0.02
Diabetes, complicated	0.02	0.07
Heart attack	0.01	0.02
Liver disease	0.01	0.02

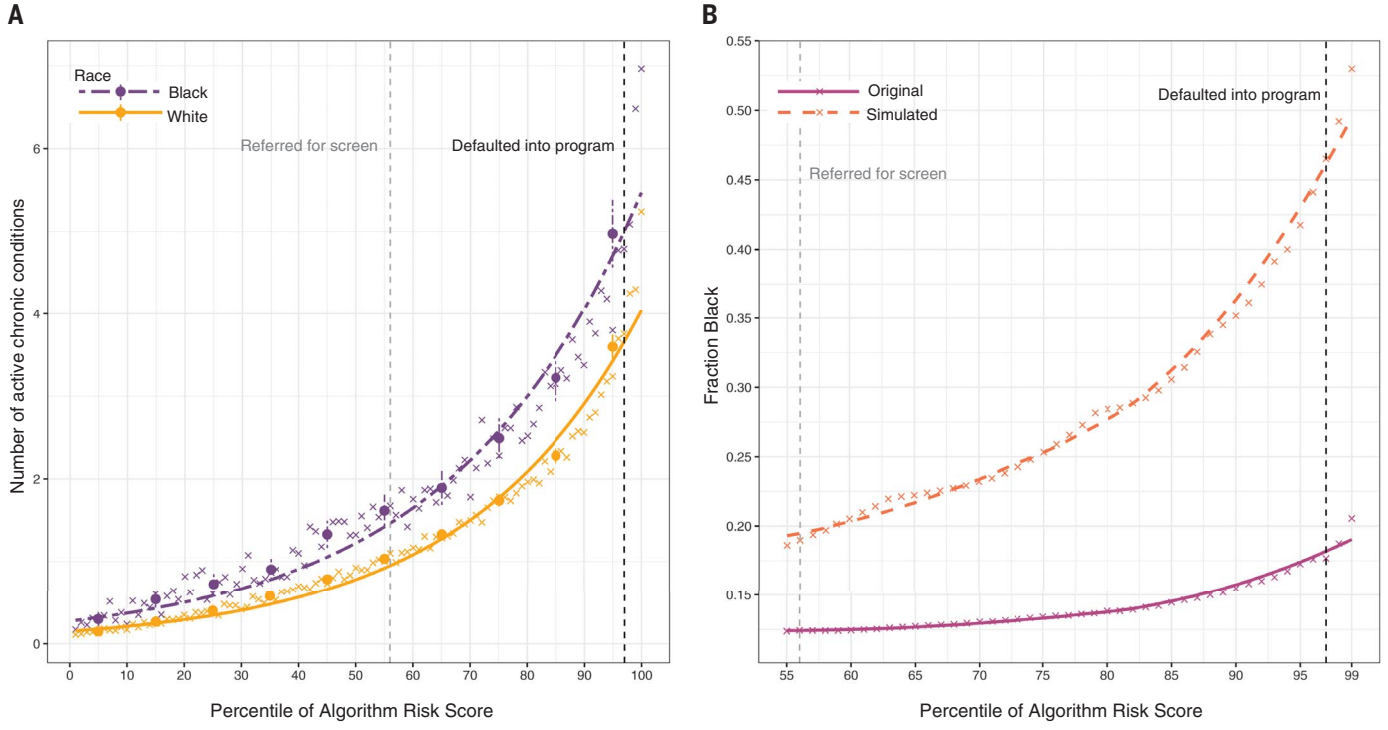

Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

of substantial disparities in program screening. We quantify this by simulating a counterfactual world with no gap in health conditional on risk. Specifically, at some risk threshold α, we identify the supramarginal White patient (i) with $R_i > \alpha$ and compare this patient's health to that of the inframarginal Black patient (j) with $R_j < \alpha$. If $H_i > H_j$, as measured by number of chronic medical conditions, we replace the (healthier, but supramarginal) White patient with the (sicker, but inframarginal) Black patient. We repeat this procedure until $H_i = H_j$, to simulate an algorithm with no predictive gap between Blacks and Whites. Fig. 1B shows the results: At all risk thresholds α above the 50th percentile, this procedure would increase the fraction of Black patients. For example, at α = 97th percentile, among those auto-identified for the program, the fraction of Black patients would rise from 17.7 to 46.5%.

We then turn to a more multidimensional picture of the complexity and severity of patients' health status, as measured by biomarkers that index the severity of the most common chronic illnesses in our sample (as shown in Table 1). This allows us to identify patients who might derive a great deal of benefit from care management programs—e.g., patients with severe diabetes who are at risk of catastrophic complications if they do not lower their blood sugar (*18*, *26*). (The materials and methods section describes several experiments to rule out a large effect of the program on these health measures in year t; had there been such an effect, we could not easily use the measures to assess the accuracy of the algorithm's predictions on health, because the program is allocated as a function of algorithm score.) Across all of these important markers of health needs—severity of diabetes, high blood pressure, renal failure, cholesterol, and anemia—we find that Blacks are substantially less healthy than Whites at any level of algorithm predictions, as shown in Fig. 2. Blacks have more-severe hypertension, diabetes, renal failure, and anemia, and higher cholesterol. The magnitudes of these differences are large: For example, differences in severity of hypertension (systolic pressure: 5.7 mmHg) and diabetes [glycated hemoglobin (HbA1c): 0.6%] imply differences in all-cause mortality of 7.6% (*27*) and 30% (*28*), respectively, calculated using data from clinical trials and longitudinal studies.

Mechanism of bias

An unusual aspect of our dataset is that we observe the algorithm's inputs and outputs as well as its objective function, providing us a unique window into the mechanisms by which bias arises. In our setting, the algorithm takes in a large set of raw insurance claims data $X_{i,t-1}$ (features) over the year $t-1$: demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs. Notably, the algorithm specifically excludes race.

The algorithm uses these data to predict $Y_{i,t}$ (i.e., the label). In this instance, the algorithm takes total medical expenditures (for simplicity, we denote "costs" C_t) in year t as the label. Thus, the algorithm's prediction on health needs is, in fact, a prediction on health costs.

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. Fig. 3A shows that, at every level of algorithm-predicted risk, Blacks and Whites have (roughly) the same costs the following year. In other words, the algorithm's predictions are well calibrated across races. For example, at the median risk score, Black patients had costs of $5147 versus $4995 for Whites (U.S. dollars); in the top 5% of algorithm-predicted risk, costs were $35,541 for Blacks versus $34,059 for Whites.

Because these programs are used to target patients with high costs, these results are largely inconsistent with algorithmic bias, as measured by calibration: Conditional on risk score, predictions do not favor Whites or Blacks anywhere in the risk distribution.

To summarize, we find substantial disparities in health conditional on risk but little disparity in costs. On the one hand, this is surprising: Health care costs and health needs are highly correlated, as sicker patients need and receive more care, on average. On the other hand, there are many opportunities for a wedge to creep in between needing health care and receiving health care—and crucially, we find that wedge to be correlated with race, as shown in Fig. 3B. At a given level of health (again measured by number of chronic illnesses), Blacks generate lower costs than Whites—on average, $1801 less per year, holding constant the number of chronic illnesses (or $1144 less, if we instead hold constant the specific individual illnesses that contribute to the sum). Table S2 also shows that Black patients generate very different kinds of costs: for example, fewer inpatient surgical and outpatient specialist costs, and more costs related to emergency visits and dialysis. 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. As a result, accurate prediction of costs necessarily means being racially biased on health.

How might these disparities in cost arise? The literature broadly suggests two main potential channels. First, poor patients face substantial barriers to accessing health care, even when enrolled in insurance plans. Although the population we study is entirely insured, there are many other mechanisms by which poverty can lead to disparities in use of health care: geography and differential access to transportation, competing demands from jobs or child care, or knowledge of reasons to seek care (*29*–*31*). To the extent that race and socioeconomic status are correlated, these factors will differentially affect Black patients. Second, race could affect costs directly via several channels: direct ("taste-based") discrimination, changes to the doctor–patient relationship, or others. A recent trial randomly assigned Black patients to a Black or White primary care provider and found significantly higher uptake of recommended preventive care when the provider was Black (*32*). This is perhaps the most rigorous demonstration of this effect, and it fits with a larger literature on potential mechanisms by which race can affect health care directly. For example, it has long been documented that Black patients have reduced trust in the health care system (*33*), a fact that some studies trace to the revelations of the Tuskegee study and other adverse experiences (*34*). A substantial literature in psychology has documented physicians' differential perceptions of Black patients, in terms of intelligence, affiliation (*35*), or pain tolerance (*36*). Thus, whether it is communication, trust, or bias, something about the interactions of Black patients with the health care system itself leads to reduced use of health care. The collective effect of these many channels is to lower health spending substantially for Black patients, conditional on need—a finding that has been appreciated for at least two decades (*37*).

Problem formulation

Our findings highlight the importance of the choice of the label on which the algorithm is trained. On the one hand, the algorithm manufacturer's choice to predict future costs is reasonable: The program's goal, at least in part, is

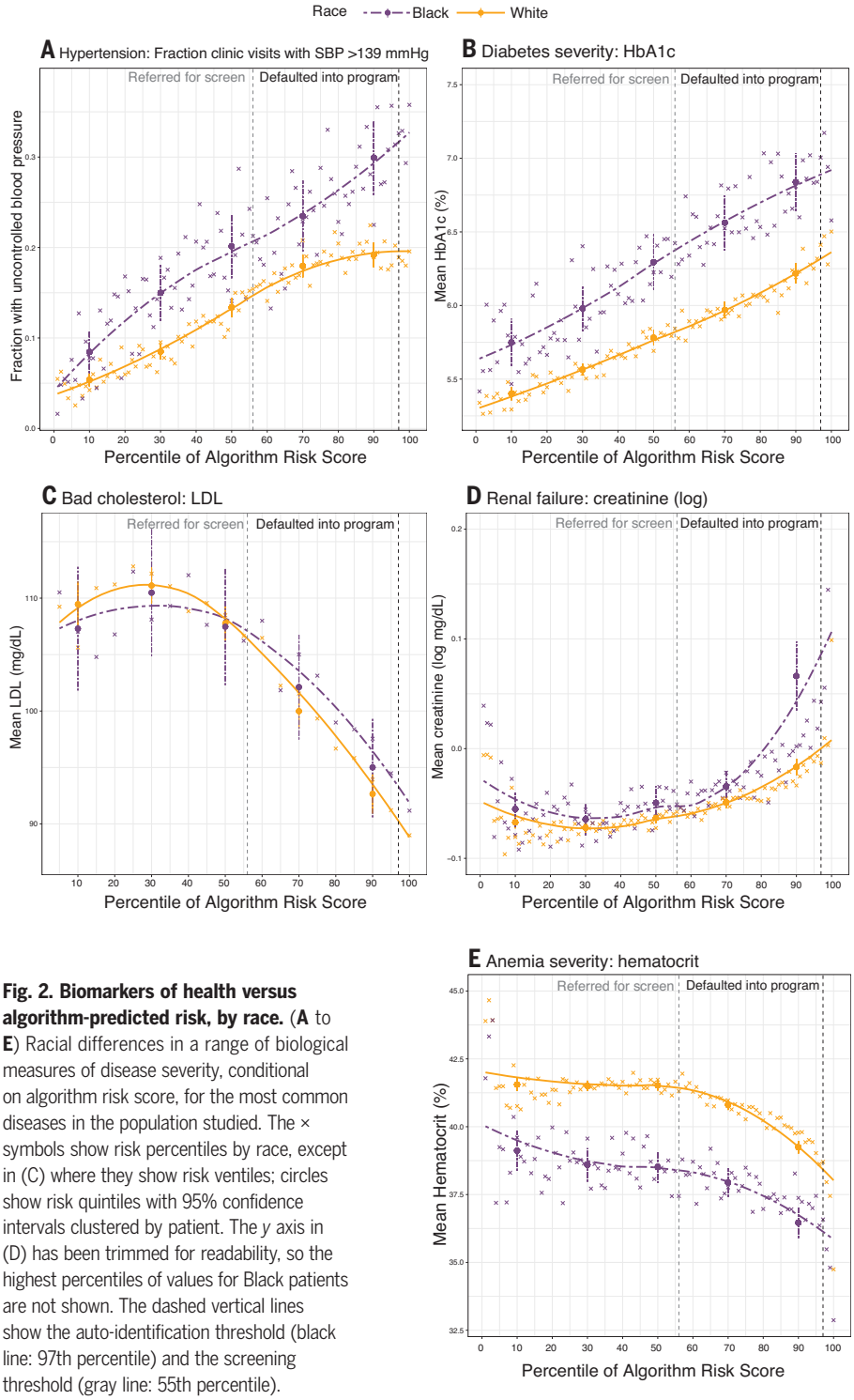

Fig. 2. Biomarkers of health versus algorithm-predicted risk, by race. (**A** to **E**) Racial differences in a range of biological measures of disease severity, conditional on algorithm risk score, for the most common diseases in the population studied. The × symbols show risk percentiles by race, except in (C) where they show risk ventiles; circles show risk quintiles with 95% confidence intervals clustered by patient. The *y* axis in (D) has been trimmed for readability, so the highest percentiles of values for Black patients are not shown. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile).

to reduce costs, and it stands to reason that patients with the greatest future costs could have the greatest benefit from the program. As noted in the supplementary materials, the manufacturer is not alone. Although the details of individual algorithms vary, the cost label reflects the industry-wide approach. For example, the Society of Actuaries's comprehensive evaluation of the 10 most widely used algorithms, including the particular algorithm we study, used cost prediction as its accuracy metric (*21*). As noted in the report, the enthusiasm for cost prediction is not restricted to industry: Similar algorithms are developed and used by non-profit hospitals, academic groups, and governmental agencies, and are often described in academic literature on targeting population health interventions (*18*, *19*).

On the other hand, future cost is by no means the only reasonable choice. For example, the evidence on care management programs shows that they do not operate to reduce costs globally. Rather, these programs primarily work to prevent acute health decompensations that lead to catastrophic health care utilization (indeed, they actually work to increase other categories of costs, such as primary care and home health assistance; see table S2). Thus avoidable future costs, i.e., those related to emergency visits and hospitalizations, could be a useful label to predict. Alternatively, rather than predicting costs at all, we could simply predict a measure of health; e.g., the number of active chronic health conditions. Because the program ultimately operates to improve the management of these conditions, patients with the most encounters related to them could also be a promising group on which to deploy preventative interventions.

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 (*38*). 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 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.

Experiments on label choice

Through a series of experiments with our dataset, we can gain some insight into how label choice affects both predictive performance and racial bias. We develop three new predictive algorithms, all trained in the same way, to predict the following outcomes: total cost in year t (this tailors cost predictions to our own dataset rather than the national training set), avoidable cost in year t (due to emergency visits and hospitalizations), and health in year t (measured by the number of chronic conditions that flare up in that year). We train all models in a random ⅔ training set and show all results only from the ⅓ holdout set. Furthermore, as with the original algorithm, we exclude race from the feature set (more details are in the materials and methods).

Table 2 shows the results of these experiments. The first finding is that all algorithms perform reasonably well for predicting not only the outcome on which they were trained but also the other outcomes: The concentration of realized outcomes in those at or above the 97th percentile is notably similar for all algorithms across all outcomes. The largest difference in performance across algorithms is seen for cost prediction: Of all costs in the holdout set, the fraction generated by those at or above the 97th percentile is 16.5% for the cost predictor versus 12.1% for the predictor

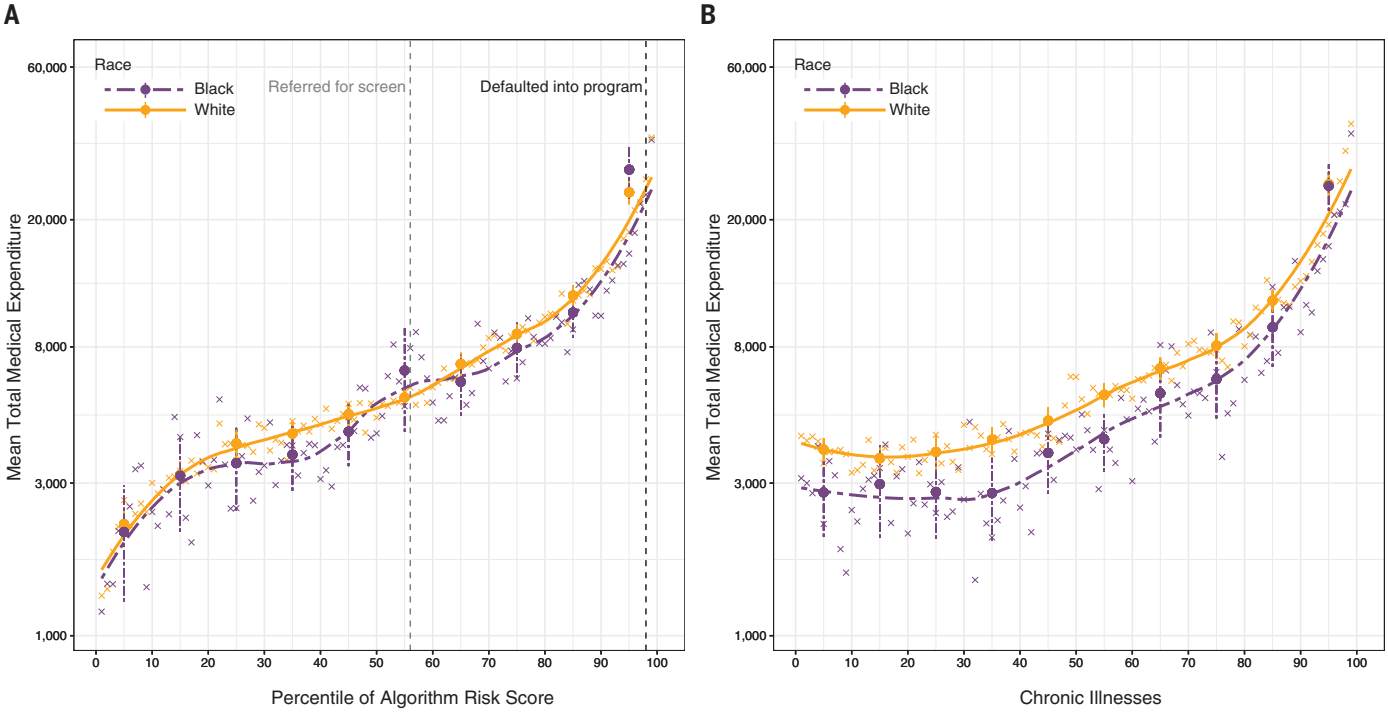

Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (**A**) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (**B**) Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The *y* axis uses a log scale.

of chronic conditions. We then test for label choice bias, defined analogously to calibration bias above: For two algorithms trained to predict Y and Y', and using a threshold τ indexing a (similarly sized) high-risk group, we would test $p[B|R > \tau] = p[B|R' > \tau]$ (here, p denotes probability and B represents Black patients).

We find that the racial composition of this highest-risk group varies far more across algorithms: The fraction of Black patients at or above these risk levels ranges from 14.1% for the cost predictor to 26.7% for the predictor of chronic conditions. Thus, although there could be many reasonable choices of label—all predictions are highly correlated, and any could be justified as a measure of patients' likely benefit from the program—they have markedly different implications in terms of bias, with nearly twofold variation in composition of Black patients in the highest-risk groups.

Relation to human judgment

As noted above, the algorithm is not used for program enrollment decisions in isolation. Rather, it is used as a screening tool, in part to alert primary care doctors to high-risk patients. 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 (for instance, primary care practice site, residence outside of a nursing home, and continual enrollment in an insurance plan).

Table 3 shows statistics on those enrolled in the program, accounting for 1.3% of observations in our sample: The enrolled individuals are 19.2% Black (versus 11.9% Black in our entire sample) and account for 2.9% of all costs and 3.3% of all active chronic conditions in the population as a whole. We then perform four counterfactual simulations to put these numbers in context; naturally, these simulations use only observable factors, not the many unobserved administrative and human factors that also affect enrollment. First, we calculate the realized program enrollment rate within each percentile of the original algorithm's predicted risk bins and randomly sample patients in each bin for enrollment. This simulation, which mimics "race-blind" enrollment conditional on algorithm score, would yield an enrolled population that is 18.3% Black (versus 19.2% observed; P = 0.8348). Second, rather than randomly sampling, we sample those with the highest predicted number of active chronic conditions within a risk bin (using our experimental algorithm described above); this would yield a population that is 26.9% Black. Finally, we compare this to simply assigning those with the highest predicted costs, or the highest number of active chronic conditions, to the program (also using our own algorithms detailed above), which would yield 17.2 and 29.2% Black patients, respectively. 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.

Discussion

Bias attributable to label choice—the difference between some unobserved optimal prediction and the prediction of an algorithm trained on an observed label—is a useful framework through which to understand bias in algorithms, both

Table 2. Performance of predictors trained on alternative labels. For each new algorithm, we show the label on which it was trained (rows) and the concentration of a given outcome of interest (columns) at or above the 97th percentile of predicted risk. We also show the fraction of Black patients in each group.

Algorithm training label	Concentration in highest-risk patients (SE)						Fraction of Black patients in group with highest risk (SE)	
	Total costs		Avoidable costs		Active chronic conditions			
Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210	(0.003)
Active chronic conditions	0.121	(0.003)	0.182	(0.003)	0.148	(0.003)	0.267	(0.003)
Best-to-worst difference	0.044		0.033		0.043		0.126	

Table 3. Doctors' decisions versus algorithmic predictions. For those enrolled in the high-risk care management program (1.3% of our sample), we first show the fraction of the population that is Black, as well as the fraction of all costs and chronic conditions accounted for by these observations. We also show these quantities for four alternative program enrollment rules, which we simulate in our dataset (using the holdout set when we use our experimental predictors). We first calculate the program enrollment rate within each percentile bin of predicted risk from the original algorithm and either (i) randomly sample patients or (ii) sample those with the highest predicted number of active chronic conditions within a bin and assign them to the program. The resultant values are then compared with values obtained by simply assigning the aforementioned 1.3% of our sample with (iii) the highest predicted cost or (iv) the highest number of active chronic conditions to the program.

Population	Fraction Black (SE)		Fraction of all costs (SE)		Fraction of all active chronic conditions (SE)	
Observed program enrollment (1.3%)	0.192	(0.003)	0.029	(0.001)	0.033	(0.001)
Simulated alternative enrollment rules						
Random, in predicted-cost bin	0.183	(0.003)	0.044	(0.002)	0.034	(0.001)
Predicted health, in predicted-cost bin	0.269	(0.003)	0.044	(0.002)	0.064	(0.002)
Highest predicted cost	0.172	(0.003)	0.100	(0.002)	0.047	(0.002)
Worst predicted health	0.292	(0.004)	0.067	(0.002)	0.076	(0.002)

in the health sector and further afield. This is because labels are often measured with errors that reflect structural inequalities (*39*). Within the health sector, using mortality or readmission rates to measure hospital performance penalizes those serving poor or non-White populations (*40*, *41*). Outside of the health arena, credit-scoring algorithms predict outcomes related to income, thus incorporating disparities in employment and salary (*2*). Policing algorithms predict measured crime, which also reflects increased scrutiny of some groups (*42*). Hiring algorithms predict employment decisions or supervisory ratings, which are affected by race and gender biases (*43*). Even retail algorithms, which set pricing for goods at the national level, penalize poorer households, which are subjected to increased prices as a result (*44*).

This mechanism of bias is particularly pernicious because it can arise from reasonable choices: Using traditional metrics of overall prediction quality, cost seemed to be an effective proxy for health yet still produced large biases. After completing the analyses described above, we contacted the algorithm manufacturer for an initial discussion of our results. In response, the manufacturer independently replicated our analyses on its national dataset of 3,695,943 commercially insured patients. This effort confirmed our results—by one measure of predictive bias calculated in their dataset, Black patients had 48,772 more active chronic conditions than White patients, conditional on risk score—illustrating how biases can indeed arise inadvertently.

To resolve the issue, we began to experiment with solutions together. As a first step, we suggested using the existing model infrastructure—sample, predictors (excluding race, as before), training process, and so forth—but changing the label: Rather than future cost, we created an index variable that combined health prediction with cost prediction. This approach reduced the number of excess active chronic conditions in Blacks, conditional on risk score, to 7758, an 84% reduction in bias. Building on these results, we are establishing an ongoing (unpaid) collaboration to convert the results of Table 3 into a better, scaled predictor of multidimensional health measures, with the goal of rolling these improvements out in a future round of algorithm development. Of course, our experience may not be typical of all algorithm developers in this sector. But because the manufacturer of the algorithm we study is widely viewed as an industry leader in data and analytics, we are hopeful that this endeavor will prompt other manufacturers to implement similar fixes.

These results suggest that label biases are fixable. Changing the procedures by which we fit algorithms (for instance, by using a new statistical technique for decorrelating predictors with race or other similar solutions) is not required. Rather, we must change the data we feed the algorithm—specifically, the labels we give it. Producing new labels requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment. But there is precedent for all of these functions in the literature and, more concretely, in the private companies that invest heavily in developing new and improved labels to predict factors such as consumer behavior (*45*). In addition, although health—as well as criminal justice, employment, and other socially important areas—presents substantial challenges to measurement, the importance of these sectors emphasizes the value of investing in such research. Because labels are the key determinant of both predictive quality and predictive bias, careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their risks.

REFERENCES AND NOTES

1. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine Bias," *ProPublica* (23 May 2016); www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
2. S. Barocas, A. D. Selbst, *Calif. Law Rev.* **104**, 671 (2016).
3. A. Chouldechova, A. Roth, arXiv:1810.08810 [cs.LG] (20 October 2018).
4. A. Datta, M. C. Tschantz, A. Datta, *Proc. Privacy Enhancing Technol.* **2015**, 92–112 (2015).
5. L. Sweeney, *Queue* **11**, 1–19 (2013).
6. M. Kay, C. Matuszek, S. A. Munson, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015), pp. 3819–3828.
7. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, A. K. Jain, *IEEE Trans. Inf. Forensics Security* **7**, 1789–1801 (2012).
8. J. Buolamwini, T. Gebru, in *Proceedings of the Conference on Fairness, Accountability and Transparency* (PMLR, 2018), pp. 77–91.
9. A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183–186 (2017).
10. S. Corbett-Davies, S. Goel, arXiv:1808.00023 [cs.CY] (31 July 2018).
11. M. De-Arteaga *et al.*, arXiv:1901.09451 [cs.IR] (27 January 2019).
12. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 259–268.
13. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, *Q. J. Econ.* **133**, 237–293 (2018).
14. C. S. Hong, A. L. Siegel, T. G. Ferris, *Issue Brief (Commonwealth Fund)* **19**, 1–19 (2014).
15. N. McCall, J. Cromwell, C. Urato, "Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH)" (RTI International, 2010).
16. J. Hsu *et al.*, *Health Aff.* **36**, 876–884 (2017).
17. L. Nelson, "Lessons from Medicare's demonstration projects on disease management and care coordination" (Working Paper 2012-01, Congressional Budget Office, 2012).
18. C. Vogeli *et al.*, *J. Gen. Intern. Med.* **22** (suppl. 3), 391–395 (2007).
19. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, *Health Aff.* **33**, 1123–1131 (2014).
20. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **105**, 491–495 (2015).
21. G. Hileman, S. Steele, "Accuracy of claims-based risk scoring models" (Society of Actuaries, 2016).
22. J. Kleinberg, S. Mullainathan, M. Raghavan, arXiv:1609.05807 [cs.LG] (19 September 2016).
23. A. Chouldechova, *Big Data* **5**, 153–163 (2017).
24. V. de Groot, H. Beckerman, G. J. Lankhorst, L. M. Bouter, *J. Clin. Epidemiol.* **56**, 221–229 (2003).
25. J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, *J. Clin. Epidemiol.* **64**, 749–759 (2011).
26. A. K. Parekh, M. B. Barton, *JAMA* **303**, 1303–1304 (2010).
27. D. Ettehad *et al.*, *Lancet* **387**, 957–967 (2016).
28. K.-T. Khaw *et al.*, *BMJ* **322**, 15 (2001).
29. K. Fiscella, P. Franks, M. R. Gold, C. M. Clancy, *JAMA* **283**, 2579–2584 (2000).
30. N. E. Adler, K. Newman, *Health Aff.* **21**, 60–76 (2002).
31. N. E. Adler, W. T. Boyce, M. A. Chesney, S. Folkman, S. L. Syme, *JAMA* **269**, 3140–3145 (1993).
32. M. Alsan, O. Garrick, G. C. Graziani, "Does diversity matter for health? Experimental evidence from Oakland" (National Bureau of Economic Research, 2018).
33. K. Armstrong, K. L. Ravenell, S. McMurphy, M. Putt, *Am. J. Public Health* **97**, 1283–1289 (2007).
34. M. Alsan, M. Wanamaker, *Q. J. Econ.* **133**, 407–455 (2018).
35. M. van Ryn, J. Burke, *Soc. Sci. Med.* **50**, 813–828 (2000).
36. K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4296–4301 (2016).
37. J. J. Escarce, F. W. Puffer, in *Racial and Ethnic Differences in the Health of Older Americans* (National Academies Press, 1997), chap. 6; www.ncbi.nlm.nih.gov/books/NBK109841/.
38. S. Passi, S. Barocas, arXiv:1901.02547 [cs.CY] (8 January 2019).
39. S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **107**, 476–480 (2017).
40. K. E. Joynt Maddox *et al.*, *Health Serv. Res.* **54**, 327–336 (2019).
41. K. E. Joynt Maddox, M. Reidhead, A. C. Qi, D. R. Nerenz, *JAMA Intern. Med.* **179**, 769–776 (2019).
42. K. Lum, W. Isaac, *Significance* **13**, 14–19 (2016).
43. I. Ajunwa, "The Paradox of Automation as Anti-Bias Intervention," available at SSRN (2016); https://ssrn.com/abstract=2746078.
44. S. DellaVigna, M. Gentzkow, "Uniform pricing in US retail chains" (National Bureau of Economic Research, 2017).
45. C. A. Gomez-Uribe, N. Hunt, *ACM Trans. Manag. Inf. Syst.* **6**, 13 (2016).

ACKNOWLEDGMENTS

We thank S. Lakhtakia, Z. Li, K. Lin, and R. Mahadeshwar for research assistance and D. Buefort and E. Maher for data science expertise. **Funding:** This work was supported by a grant from the National Institute for Health Care Management Foundation. **Author contributions:** Z.O. and S.M. designed the study, obtained funding, and conducted the analyses. All authors contributed to reviewing findings and writing the manuscript. **Competing interests:** The analysis was completely independent: None of the authors had any contact with the algorithm's manufacturer until after it was complete. No authors received compensation, in any form, from the manufacturer or have any commercial interests in the manufacturer or competing entities or products. There were no confidentiality agreements that limited reporting of the work or its results, no material transfer agreements, no oversight in the preparation of this article (besides ethical oversight from the approving IRB, which was based at a non-profit academic health system), and no formal relationship of any kind between any of the authors and the manufacturer. **Data and materials availability:** Because the data used in this analysis are protected health information, they cannot be made publicly available. We provide instead a synthetic dataset (using the R package synthpop) and all code necessary to reproduce our analyses at https://gitlab.com/labsysmed/dissecting-bias.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/366/6464/447/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S4
References (*46*–*51*)

8 March 2019; accepted 4 October 2019
10.1126/science.aax2342

Science

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer, Brian Powers, Christine Vogeli and Sendhil Mullainathan

Science **366** (6464), 447-453.
DOI: 10.1126/science.aax2342

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Science, this issue p. 447; see also p. 421

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.