Stanford University
Human-Centered
Artificial Intelligence

OCTOBER
2020

SPECIAL HEALTH CARE SERIES

Toward Fairness in Health Care Training Data

Amit Kaushal, Russ Altman and Curt Langlotz

With recent advances in artificial intelligence (AI), researchers can now train sophisticated computer algorithms to interpret medical images – often with accuracy comparable to trained physicians. Yet our recent survey of medical research shows that these algorithms rely on datasets that lack population diversity and could introduce bias into the understanding of a patient's health condition.

Artificial intelligence algorithms increasingly inform the decisions of human experts. In medical imaging, these algorithms may help a doctor spot a subtle finding or suggest an alternate diagnosis. But bias in the data used to train these high-stakes algorithms can bias the algorithm itself. Our analysis shows that the datasets used to develop these algorithms come from only a handful of locations – raising serious questions for policymakers—but also providing opportunities for course correction.

In our research, published in the *Journal of the American Medical Association*, we looked at data from more than 70 studies that used U.S. patient data to train algorithms designed to compete or collaborate with physicians to perform diagnostic tasks. Overwhelmingly, the datasets came from three states—California, Massachusetts, and New York—with little or no representation from the remaining 47 states. Rectifying this lack of representation in medical data should be front of mind for health policymakers and regulators. Lack of data diversity can be addressed in part by initiatives to streamline the nation's digital infrastructure, to enhance the availability of patient data from underrepresented populations for larger studies, and to incentivize ethical data sharing and the democratization of medical data.

KEY TAKEAWAYS

- Bias arises when we build algorithms using datasets that don't mirror the population. When generalized to larger swathes of the population, these nonrepresentative data have the potential to confound research findings.
- The vast majority of the health data used to build AI algorithms came from only three states, with little or no representation from the remaining 47 states.
- Policymakers, regulators, industry, and academia need to work together to ensure medical AI data reflect America's diversity across not only geography but also many other important attributes. To that end, nationwide data sharing initiatives should be a top priority.

Introduction

Advances in computer vision and deep learning have produced algorithms that perform image-based diagnostic tasks in fields like radiology, ophthalmology, dermatology, pathology, gastroenterology, and cardiology with accuracy approaching or exceeding that of trained physicians. Despite their well-documented successes, these machine learning algorithms are vulnerable to biases when an insufficient quantity or diversity of data are used to train them.

We investigated an understudied source of systemic bias in clinical applications of deep learning: the geographic distribution of patient cohorts used to train algorithms. In our research, we examined peer-reviewed articles in a major biomedical and life sciences database (PubMed) published between January 1, 2015, and December 31, 2019. Among studies where geographic origin could be characterized, we found that algorithms employing U.S. patient data were disproportionately based on cohorts of people from California, Massachusetts, and New York, with little to no representation from the remaining 47 states.

Research Findings

Of the 74 studies that met inclusion criteria for our analysis, 56 of them trained algorithms using at least one geographically identifiable cohort. Cohorts from California appeared in 22 of the 56 studies, cohorts from Massachusetts in 15, and cohorts from New York

US PATIENT COHORTS USED FOR TRAINING CLINICAL MACHINE LEARNING ALGORITHMS, BY STATE[a]

States	No. of studies
California	22
Massachusetts	15
New York	14
Pennsylvania	5
Maryland	4
Colorado	2
Connecticut	2
New Hampshire	2
North Carolina	2
Indiana	1
Michigan	1
Minnesota	1
Ohio	1
Texas	1
Vermont	1
Wisconsin	1

[a]Fifty-six studies used 1 or more geograrphically identifiable US patient cohorts in the training of their clinical machine learning algorithm. Thirty-four states were not represented in geographically identifiable cohorts: Alabama, Alaska, Arizona, Arkansas, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Iowa, Kansas, Kentucky, Louisiana, Maine, Mississippi, Missouri, Montana, Nebraska, Nevada, New Jersey, New Mexico, North Dakota, Oklahoma, Oregon, Rhode Island, South Carolina, South Dakota, Tennessee, Utah, Virginia, Washington, West Virginia, and Wyoming.

in 14. Across all studies only 23 cohorts involving multiple states were identified. Only eighteen of the 74 studies (24%) utilized cohorts from more than one state exclusively and 34 states did not contribute any patient cohorts whatsoever. Cohorts involving multiple states were largely drawn from studies or consortia associated with the National Institutes of Health (NIH), industry trials or databases, a small number of online image atlases, and one from an academic second opinion service.

This lack of diversity is alarming because geography correlates to a number of factors ranging from lifestyle and diet to weather, exposure to dangerous chemicals, and a host of other unknown variables. This in turn carries the potential to confound the validity of some of the most promising applications of medical AI unless corrective measures are taken. Patients from California, Massachusetts, and New York have economic, educational, social, behavioral, ethnic and cultural features that are not representative of the entire nation. As a result, algorithms trained primarily on patient data from these states may not be appropriate for patients from other regions of the country. This creates risk when generalizing these diagnostic algorithms to new populations or different geographies.

Medical data collection today is hindered by a health care infrastructure that was not designed to handle the vast quantities of data needed.

Discussion

As AI starts to transform more areas of medical discovery and health care delivery, the focus should be on how it improves care and yields better outcomes. Yet medical data collection today is hindered by a health care infrastructure that was not designed to handle the vast quantities of data needed to ensure the risks discussed above are mitigated. We need to build a data infrastructure that broadly captures and makes available diverse patient data so we can better diagnose and treat people for their health challenges.

The issue is profoundly timely, as promises of artificial intelligence revolutionizing biomedicine are ubiquitous.

The issue is profoundly timely, as promises of artificial intelligence revolutionizing biomedicine are ubiquitous. It often seems we're close to launching AI systems that can remotely identify a person about to get sick, make a medical diagnosis without the doctor present, or select a custom AI-designed pharmaceutical and deliver it to the patient just in time. If indeed this is the future, we are far from reaching it unless we start collecting the most inclusive patient data possible.

What is needed now are more specific regulatory requirements for medical research and a streamlined set of procedures for building and maintaining repositories of medical data from across the country. Consistent with various proposed national visions for AI that prioritize interdisciplinary collaboration and inclusivity, a range of measures are available to policymakers hoping to reduce bias in medical AI tools.

Federal policymakers should build on a commitment to making more representative medical data available to train AI tools by continuing to champion the dissemination of diagnostic datasets and by serving as leaders in building the necessary infrastructure to make medical data truly democratic across the nation. To that end, proposals to overhaul our digital infrastructure and create data repositories should make it easier to assemble, process, and work with patient data from underrepresented populations and reduce the barriers to incorporating them into larger studies. In combination with high-end computational resources and vital expertise from academia and industry, providing these large-scale federally-supported data sets to researchers will ensure that advances in medical AI will benefit all members of society.

Legislation to democratize medical AI development across the U.S. could empower a new generation of ethical AI scientists with the tools to improve health for all. Healthcare costs could be greatly diminished and innovation supercharged if government can agree on a way to ethically aggregate and allow research access to data from the many large and small health facilities throughout our nation.

Legislation to democratize medical AI development across the U.S. could provide a counterweight to the current imbalance emphasizing industry in the innovation ecosystem.

Additional measures policymakers could consider include regional and nationwide consortia to incentivize collaboration and data sharing as well as expanded support to existing programs like the VA's National Artificial Intelligence Institute and Centers for Medicare and Medicaid's Artificial Intelligence Health Outcomes Challenge, and the health component of the Department of Defense's Joint Artificial Intelligence Center. Supporting pilot data sharing programs at the local, state, federal, and tribal government levels throughout the nation could further help us get the diverse datasets we need to transcend our present impasse.

Safeguarding patient privacy and confidentiality must remain at the forefront of these discussions. But keeping in mind the improvements in health care outcomes that these measures promise should make the prospect of their adoption attractive to partners throughout academia, industry, and government alike.

The original report, **"Geographic Distribution of U.S. Cohorts Used to Train Deep Learning Algorithms,"** can be found here: https://jamanetwork.com/journals/jama/article-abstract/2770833.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. For further information, please contact **HAI-Policy@stanford.edu.**

Amit Kaushal is a Clinical Assistant Professor of Medicine and Adjunct Professor of Bioengineering at Stanford University. His research spans clinical medicine, teaching, research, and industry.

Russ Altman is a the Kenneth Fong Professor of Bioengineering, Genetics, Medicine, Biomedical Data Science and (by courtesy) Computer Science) and past chairman of the Bioengineering Department at Stanford University. He is an associate director at Stanford's Institute for Human-Centered Artificial Intelligence (HAI).

Curtis Langlotz is a Professor of Radiology and Biomedical Informatics and Director of the Center for Artificial Intelligence in Medicine and Imaging at Stanford University. He is a faculty member at Stanford's Institute for Human-Centered Artificial Intelligence (HAI).

HAI Stanford University Human-Centered Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**