Health Data Platforms

Sendhil Mullainathan, University of Chicago and NBER
Ziad Obermeyer, UC Berkeley and NBER[1]

Abstract

Despite great optimism for the role that artificial intelligence may play in health care, actual applications of AI to health are in short supply. Using a concrete application of AI to clinical decision-making, we argue that the bottleneck is not financial or behavioral, but the scarcity of usable clinical data. Health data platforms that allow safe and ethical access to health data could catalyze both research and development in this area.

Introduction

Clinical medicine is ripe for transformation by artificial intelligence. The field draws a wide range of highly intelligent people, motivated both by social impact and by the chance to profit from an industry that accounts for 20% of US GDP. The health care system also has myriad known inefficiencies in human decision making, the correction of which could create enormous social value. And it produces exabytes of data every year that could easily be used to train useful algorithms.

What then accounts for the conspicuous lack of AI deployed in clinics and hospitals today? Commonly cited factors for this—and many other problems in the health care system—are misaligned financial incentives, regulatory barriers, and behavioral factors. While we are not blind to the challenges these factors present, here we argue for another: the lack of accessible clinical data. We begin by discussing a concrete use case for clinical AI: a tool to help physicians test for heart attack (acute coronary syndromes) in the emergency setting. We then discuss the major barrier to implementing this tool: data. Finally, we make the case for new mechanisms that connect researchers and clinical product developers to health data.

Testing for heart attack in the emergency department (ED)

In recent work, we study how physicians arrive at an important diagnostic decision: whether or not to test a patient for heart attack, new blockages in the coronary arteries supplying the heart (Mullainathan and Obermeyer, 2022a). Patients with heart attack need urgent treatment: untreated blockages can cause heart failure and chronic pain (angina), or arrythmia and sudden death. Blockages cannot be diagnosed from symptoms alone. For example, the most common symptom, chest pain, could also be the result of a pinched nerve, or acid reflux, for example. And while the laboratory tests and electrocardiograms done in the ED can be suggestive of blockage, they are not conclusive. As such, physicians rely on an invasive procedure, cardiac catheterization, which can be performed either directly or after a lower-cost 'stress test.' Since

[1] Both authors are co-founders and hold equity in Dandelion Health, a for-profit entity discussed in this paper.

these tests are costly, they are applied only when the probability of a heart attack is judged to be sufficiently high.

Central to our approach is a set of tools from the field of machine learning. Because the value of a test depends on the likelihood it will come back positive, the testing decision can be cast as a 'prediction problem,' where machine learning techniques can be profitably applied (Agrawal et al., 2018; Kleinberg et al., 2015). In deciding whom to test, the physician must in effect form a prediction. A test sure to come back negative is a waste; and except at the extreme, the value of a test increases with the probability it will come back positive. As such, efficient testing is grounded in effective predicting. By providing explicit predictions, algorithms provide a natural benchmark against which actual decisions can be contrasted. We thus view testing for heart attack as both an important problem in its own right, and as a 'model system' for applying machine learning to study diagnostic judgments more generally.

We implement our approach on electronic health record data from a large academic medical center. A subset of these is used to train an ensemble machine learning model, that uses thousands of variables to predict whether a given test will reveal a coronary blockage. In addition, we build a cost effectiveness model that, for a set of patients at a given predicted risk, calculates the implied cost of a life year for testing that set, based on known benefits and costs of treatment. To check the generality of our results, we replicate them in a 20% nationwide sample of emergency visits by Medicare beneficiaries from. These data, based on insurance claims, are less detailed. But because they are nationally representative, allow us to explore the relevance of our results for health policy.

We first examine those patients whom physicians choose to test. Our strategy is to use the algorithm's risk predictions to identify *potentially* low-risk patients, in whom testing might not be useful. We then look at *realized* test results to see who was right—the algorithm or the physician. This allows us to calculate the actual value of testing ex post, as a function of predicted risk ex ante, and identify predictably low-value testing. The value of testing is typically adjudicated on the basis of its average yield, and in our setting, the average test has an implied cost effectiveness of $89,714 per life year. At typical US life-year valuations of $100-150,000 (Neumann et al., 2014), this would be considered cost-effective. But this aggregate statistic hides a great deal of highly inefficient testing: binning patients by predicted risk reveals that, at a threshold of $150,000 per life year, 62% of tests should be cut based on predicted risk. The bottom bin of tests is extremely cost-ineffective: $1,352,466 per life- year. For comparison, biologics for rare diseases (some of the least cost-effective technologies that health systems sometimes pay for) are typically estimated at around $300,000 per quality adjusted life-year. Even the second-lowest bin is very cost-ineffective at $318,603 dollars per life-year. By contrast, in the highest-risk quintile bins, tests cost only $46,017 per life-year, comparable with cost-effective interventions like dialysis.

The existence of over-testing is not surprising. But as an emerging economics literature suggests, it often coexists with under-testing. In particular, doctors both over-treat low risk patients, and under-treat high risk patients, in a way that suggests errors in diagnostic judgment, and under-use may be at least as consequential as over-use, as has been previously suggested (Currie and MacLeod, 2017). In our data, we find that among patients in the highest decile of predicted risk,

where testing would appear to be very cost-effective, only 38.3% are actually tested. This fact raises the possibility of under-testing—but does not fully establish it. The key econometric problem is that we do not know what *would have* happened if we tested these patients.

To answer this question, we look to new data, on 'major adverse cardiac events' in the 30 days after the patient's ER visit, that suggest undiagnosed and untreated heart attack. We find that in the highest-risk bin, untested patients go on to have an adverse cardiac event rate of 15.6%, high enough for clinical guidelines to conclude they should have been tested. We also leverage exogenous variation in who is tested—some ED shifts test patients more than others—to simulate an experiment, in which a patient is more or less likely to be tested, for reasons unrelated to their actual risk. This shows that the highest-risk patients—and only the highest-ris patients—who arrive during the highest-testing shifts have significantly lower mortality (2.5 percentage points, or 32%). In contrast, when we look on average, across all patients, without the benefit of machine learning predictions, there is no effect: increasing testing has no statistically significant effect on health outcomes, matching what is often called 'flat of the curve' health care.

Barriers to implementation of AI tools

By identifying ex ante which patients ought not to be tested, algorithmic predictions pave the way for targeted interventions to increase the efficiency of testing prospectively. Our current efforts to implement this algorithm as a clinical decision aid have given us insights into why such algorithms are not more widespread.

It is first important to note which barriers are unlikely to pose important barriers. As our results on over- and under-testing above illustrate, such an algorithm would be of great interest to health systems irrespective of their financial incentives. Consider the case of a purely profit-motivated health system that placed no weight on improved patient outcomes. If such a system were paid under a traditional fee-for-service plan, they would be highly incentivized to implement the algorithm to increase testing of high-risk patients: these patients are highly profitable, because they are the most likely to generate the complex procedures and intensive care needs that are major contributors to hospitals' bottom lines (Abelson and Creswell, 2012); they are certainly more profitable than a negative test. If by contrast such a system were paid under a risk-based (capitated) model, they too would be highly incentivized to implement the algorithm to reduce testing of low risk patients: under widely-accepted cost-effectiveness rules, about two-thirds of all tests could be cut using information available at the time of the physician's decision.

Regulatory factors are another commonly-cited reason that clinical AI is not more widely deployed. But, as recent scholarship demonstrates, the FDA's regulatory approach cannot be a binding constraint: it has approved hundreds of software and AI devices over the past decade (Stern, 2022). Finally, behavioral barriers to adoption are widely believed to be common. But in our ongoing experience designing the roll-out of the tool in a large hospital system in the setting of a large-scale randomized trial, physicians are eager to adopt tools that improve their own performance. The important caveat is that they must first be convinced, in a data-driven way, that the tools will help. We are addressing these very reasonable concerns by performing in-depth

review of individual cases, particularly those with poor patient outcomes, where the algorithm's predictions might have helped.

Rather, we believe the most significant barrier to development and implementation of such algorithms is data. It is instructive to consider how we gained access to the data we needed to build the original algorithm described above. Building up the dataset took years of effort: to identify where the data were housed, clean it and extract it. Creating the data frame itself—establishing the sample and exclusion criteria, creating the key study outcomes, developing a data-driven definition for missed heart attack, etc.—demanded both deep medical knowledge and careful applied microeconomic data work. Merging in the electrocardiogram (ECG) waveforms was particularly challenging: after identifying the source system and securing access, we discovered the waveforms were stored only as a PDF image file. So we had to write the code to extract the numeric time series from the image.

Even this painstaking work depended on solving a host of logistical issues regarding data access, which are difficult for many researchers to overcome. The electronic health record data was accessible only because one of us was, at the time, an employee of the academic hospital from which the data were sourced. This arrangement is the norm: data are completely inaccessible to those who do not have the good fortune to be administratively based within a given hospital—even faculty members at universities affiliated with the hospital are typically ineligible, meaning that economists or computer scientists wishing to access data at their university's hospital must first identify a collaborator employed by that hospital. For example, at the time we began the work for the project described above, we were both faculty members at the same university, but only one of us had access to the data—an idiosyncratic result of being on paper an employee of one of the university's affiliated hospitals. Any analyst we wished to use to assist on the project needed to be hired as an employee of the hospital system, meaning that grant funding needed to be obtained and housed at the hospital in question. Taken together, these restrictions provide a major obstacle to cross-disciplinary work, and were a major reason that the work underlying the paper took over 8 years to complete.[2]

The need for health data platforms

Scientific fields need data to grow and thrive. In economics, the availability of stock market data created the field of quantitative finance; the Medicare claims data hosted at the National Bureau of Economic Research are the foundation of our current understanding of the health care system. In computer science, datasets like DARPA's early efforts with Canadian Hansards data, to the recent examples of the Netflix Prize, MNIST, ImageNet, LFW, 1 billion words, and others underlie unprecedented recent progress in translation, sentiment analysis, object and facial recognition, and other tasks (Donoho, 2017).

In medicine, by contrast, well-connected researchers are lucky enough to have access to health data by virtue of their employment status or personal connections. Access for everyone else is laborious, costly, time-consuming, or just impossible. But it is clearly inefficient for only a small group of in-house researchers to have access to data: in addition to a simple numbers game,

[2] An honest accounting of the project timeline would place a still larger share of the blame on both authors.

where discoveries are more likely to be made if more researchers are taking shots on goal, there is also misallocation: in machine learning in particular, hospitals are unlikely to win the war for talent when competing with the deep resources of technology companies. And researchers based at well-resourced academic hospitals are likely to work on problems that concern them and their patient populations, while the needs of other populations may be ignored (Kaushal et al., 2020).

A commonly-cited concern is the protection of patient privacy. But given the many technical solutions to this problem, from sophisticated deidentification methods to highly secure cloud environments, this cannot be the only barrier. Rather, we believe the problem is incentives. Open data are a classic public good: market forces do not favor their creation. While they have enormous benefit to everyone in the long-run—patients, health systems, industry—no single actor has a strong incentive to act (Hill et al., 2020; Price and Cohen, 2019).

Emerging solutions: A data platform for academic research

We close by highlighting two health data platforms on which we have worked. The first, Nightingale Open Science, is a non-profit platform to catalyze research (Mullainathan and Obermeyer, 2022b). Thanks to philanthropic funding, Nightingale supports the creation of previously unseen datasets, in collaboration with health systems around the world. It then makes the de-identified datasets available to a diverse, global community of researchers on a secure cloud platform. By focusing on data that link medical images with real patient outcomes, the platform aims to foster groundbreaking research into common tasks at the intersection of computation and medicine.

Nightingale's datasets focus on medical imaging data: electrocardiograms, x-rays and CT scans, digital pathology images, etc. Medical images are rich sources of signal about patient health—so rich that doctors are unlikely to make full use of all the information. By contrast, most electronic health record data (e.g., diagnoses, procedures, text-based notes) are actually produced by doctors, and thus more likely to be used effectively. Standardization of imaging protocols across time and place mean that a chest x-ray in India looks much like an x-ray in San Francisco. While there is of course some variation across sites and equipment manufacturers, this is small compared to the practice and system-level variation that affects how diagnoses and other data are captured. Technical tools and legal frameworks for deidentification of medical images exist (e.g., HIPAA in the US, and many other countries' legal frameworks permit sharing). Different types of imaging present different challenges—an ECG is a simple numeric time series, while a head MRI could allow facial reconstruction—but these challenges are increasingly tractable with a robust set of tools.

These datasets were built collaboratively with a range of health systems from around the world. Diversity of data is a key consideration, given the non-representative nature of many current datasets used to build algorithms. In the San Francisco Bay Area, for example, Nightingale partners with a leading academic medical center, and also a far less well-resourced county hospital system. Abroad, partners include the largest hospital in Taiwan, and will soon expand to partnerships in Cameroon and Tamil Nadu.

Emerging solutions: A data platform for AI product development

The second platform, Dandelion Health, is a for-profit platform to catalyze AI product development. It is perhaps surprising that market forces have not solved the problem of data access in the private sector, given the large financial incentives to build AI products. While there are several different types of efforts underway to apply AI to medical datasets, these efforts are limited in several ways. First, several consortia have formed to pool electronic health record data across large hospital systems. But the center of gravity of those efforts is providing insights and analytics to life sciences companies, in the hopes of capturing a share of the large budgets associated with drug development, rather than AI product development. As a result, high-dimensional imaging and waveform data are typically absent, and complex questions about ownership of intellectual property derived from the data remain unresolved. Second, academic medical centers are beginning to partner with companies, or in some cases spinning off new ventures themselves. But these ventures are in practice limited by the complex, laborious approach to contracting, intellectual property, and data access. In addition, the unusual and non-representative nature of both the populations served by tertiary and quaternary centers, and the care practices in those centers, hampers generalizability. Third, technology companies are beginning to invest heavily in medical data, but of course their goal is to monopolize it for their own purposes, rather than to let market forces accelerate broad product development.

We do not mean to argue that such efforts are doomed to failure: many of them have the potential to produce exciting, innovative tools to improve clinical care. But they are unlikely to unlock the market forces that typically drive innovation in other sectors at scale, either because their focus is not on AI, or because data access remains limited by design.

The goal of Dandelion Health is to create the largest and highest-quality AI-ready training dataset in the world, and to become the first end-to-end product development platform for clinical AI. Dandelion has agreements with up to five massive US health systems, that allow for access to the universe of clinically-generated data: structured electronic health record data (including labs, vital signs, insurance claims data, etc.) but also notes, radiology and pathology images, neurology and cardiology waveforms, etc.

A key challenge in building up these datasets is the complexity of storage and retrieval of high-dimensional data at health systems. While tabular electronic health record data have been largely standardized thanks to large vendors, imaging and waveform data are another story entirely. The data are scattered across multiple different vendor storage systems, and hospitals are often restricted by contracts with vendors that impose high per-image costs to retrieve the hospitals' own data. More frustrating still, many hospitals continue to delete or overwrite data because of perceived storage space constraints. If data is the new oil, health systems are actively lighting one of their most precious resources on fire. Solving these problems has given us new insights into why efforts to build up datasets for clinical AI are not more widespread.

After creating the datasets within the partner health systems' environments, Dandelion deidentifies and tokenizes these data, then aggregates and curates data to support the development of new products by third parties. The goal is to securely and ethically realize the

value locked in health data—rather than letting them sit unused on health systems' servers—and use them to drive better health for patients.

Naturally, the use of patient data for product development raises legal as well as ethical issues. We have found it useful to start with the Belmont principles, which are the foundation of the ethical practice of research, while considering the complex tradeoffs in this area. The Belmont principles mandate protection of patient privacy, beneficence—in other words, doing more good than harm—and justice. These broad principles provided a clear basis for articulating the upside of data sharing for patients, while ensuring respect for their privacy and equity considerations. We feel that a similar cost-benefit tradeoff should be constantly weighed with respect to product development.

To maximize benefits, Dandelion's focus is to improve patient outcomes. As a result, the platform exists exclusively for AI innovators to create solutions that will improve patient care. Another key principle is that products made using Dandelion data should strive to reduce inequities — not exaggerate them. Dandelion conducts its own internal review to ensure that any use cases conform to this high standard, and health system partners review every client request to ensure that the products in question will actually help patients and providers. To minimize costs and risks, Dandelion protects patients' identities and privacy by de-identifying and tokenizing data before it leaves the health systems' environment. Where possible, it goes beyond existing laws to uphold the highest privacy standards. Dandelion does not own the data collected by its health system partners — nor does it sell this data to customers. Dandelion leases access to de-identified data within an encrypted SOC2-certified cloud environment.

Conclusions

Access to health data is a major bottleneck to progress in clinical AI. By investing in health data platforms to catalyze both research and development, society can realize the huge gains from AI in health that have been long promised but, to date, not delivered.

References

Abelson, R., Creswell, J., 2012. Hospital chain inquiry cited unnecessary cardiac work. New York Times 6.

Agrawal, A., Gans, J., Goldfarb, A., 2018. Prediction machines: the simple economics of artificial intelligence. Harvard Business Press.

Currie, J., MacLeod, W.B., 2017. Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians. J Labor Econ 35, 18977. https://doi.org/10.3386/w18977

Donoho, D., 2017. 50 Years of Data Science. Journal of Computational and Graphical Statistics 26, 745–766. https://doi.org/10.1080/10618600.2017.1384734

Hill, R., Stein, C., Williams, H., 2020. Internalizing Externalities: Designing Effective Data Policies. AEA Papers and Proceedings 110, 49–54. https://doi.org/10.1257/pandp.20201060

Kaushal, A., Altman, R., Langlotz, C., 2020. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA 324, 1212–1213. https://doi.org/10.1001/jama.2020.12067

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction Policy Problems. American Economic Review 105, 491–95.

Mullainathan, S., Obermeyer, Z., 2022a. Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. Quarterly Journal of Economics 137, 1–51.

Mullainathan, S., Obermeyer, Z., 2022b. Solving medicine's data bottleneck: Nightingale Open Science. Nat Med 28, 897–899. https://doi.org/10.1038/s41591-022-01804-4

Neumann, P.J., Cohen, J.T., Weinstein, M.C., 2014. Updating Cost-Effectiveness — The Curious Resilience of the $50,000-per-QALY Threshold. New England Journal of Medicine 371, 796–797. https://doi.org/10.1056/NEJMp1405158

Price, W.N., Cohen, I.G., 2019. Privacy in the age of medical big data. Nat Med 25, 37–43. https://doi.org/10.1038/s41591-018-0272-7

Stern, A.D., 2022. The Regulation of Medical AI: Policy Approaches, Data, and Innovation Incentives. Working Paper Series. https://doi.org/10.3386/w30639