Solving medicine's data bottleneck: Nightingale Open Science

Open datasets, curated around unsolved medical problems, are vital to the development of computational research in medicine, but remain in short supply. Nightingale Open Science, a non-profit computing platform, was founded to catalyse research in this nascent field.

Sendhil Mullainathan and Ziad Obermeyer

Medicine has made enormous strides in understanding how the body works, and how it fails. But deep and unsolved mysteries remain. Sudden cardiac death kills 350,000 people in the USA every year, but even in the rear-view mirror, doctors find no identifiable cause for the vast majority[1]. Cancer kills 600,000 patients in the USA every year, despite screening programs that expose millions of patients to costly, invasive tests[2]. COVID-19 has killed nearly one million people in the USA, but we still have little idea why some people die whereas others develop a runny nose or no symptoms at all[3].

Computational methods hold great promise for solving these and many other problems in medicine. Algorithms have new ways of 'seeing' patterns in the complex, high-dimensional data that health systems produce every day — electrocardiograms (ECGs), X-rays and computerized tomography (CT) scans, digital pathology images and so on —and are already yielding promising results[4–7]. Unfortunately, a major bottleneck risks stifling progress in this nascent field before it begins: the acute shortage of data accessible to researchers.

Nightingale Open Science is a computing platform designed to help to address this critical data bottleneck. Nightingale hosts massive new medical imaging datasets, curated around unsolved medical problems for which modern computational methods could be transformative. To do so, Nightingale works with health systems around the world to build datasets with two ingredients: large samples of medical images, linked to ground-truth patient outcomes. Deidentified versions of those datasets are then made available on a secure cloud platform to a diverse, global community of researchers. Thanks to a coalition of funders — anchored by Schmidt Futures, the Gordon and Betty Moore Foundation, and philanthropist Ken Griffin — the platform was launched at the NeurIPS conference in December 2021.

Table 1 | Canonical common task datasets in machine learning

	Description	Common task example
Canadian Hansards	English–French transcript of Canadian Parliamentary Debates	Machine translation
ImageNet	Images labeled with brief descriptions	Object and scene detection
Internet Movie Database	Text of consumer movie reviews labeled with quantitative ratings	Sentiment analysis
Labeled Faces in the Wild	Facial photographs with individuals indexed across photographs	Facial recognition
MNIST	Images of handwritten digits	Digit recognition
Netflix Prize	Individual user movie ratings	Recommender system
One Billion Words	Text scraped from online sources	Language modeling

Nightingale solves two problems

Like any scientific field, medicine needs data to grow and thrive. But not just any data will do. Instead, recent successes from other disciplines — genomics, computational biology, language modeling, and image recognition, to name a few — suggest that datasets must also possess two specific features. First, they must be open access: they cannot be monopolized by those who produce it, whether academics, non-profits or corporations. Instead, the data must be accessible at low cost, in terms of money and in terms of time. Only then can good ideas thrive, on a level and just playing field. Second, the data must be curated around 'common tasks': important, field-defining problems on which a community of researchers can collaborate, compete and improve. Datasets meeting these two criteria are the 'secret sauce' of machine learning — more than just computing power, or individual genius — and underlie the unprecedented recent progress in translation, sentiment analysis, object and facial recognition, and other tasks[8] (Table 1).

Existing health datasets seldom meet these two criteria. First, they are not truly open. Instead, they are often controlled by a handful of researchers at well-resourced institutions or companies. Access for everyone else is laborious, costly, time-consuming or just impossible, despite the fact that the creation of nearly all health data, whether from insurance premiums or research grants, is publicly funded. This has a variety of negative consequences. Algorithms are designed largely to serve the needs of the privileged[9]. Their performance cannot be adequately scrutinized, leading to failures of replication and erosion of trust[10]. Highly talented researchers who could make major contributions to medicine are diverted into solving trivial problems in other fields.

A commonly cited reason for these barriers to access 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 reason. Rather, 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 and industry — no single actor has a strong incentive to act (for a thoughtful review, see ref. [11]).

Second, health datasets are seldom curated in a way that allows researchers to meaningfully engage with critical questions. Specifically, they are not labeled with the

Box 1 | Why focus on imaging data?

First, medical images are rich sources of signal about patient health — so rich that doctors are unlikely to make full use of all the information contained within them. By contrast, most electronic health record data (for example, diagnoses, procedures and text-based notes) are directly produced by doctors, who are necessarily aware of the information they contain.

Second, standardization of imaging protocols across time and place means that a chest X-ray in India looks much like an X-ray in San Francisco, USA. Of course, there is some variation across sites and equipment manufacturers, but it is small compared to the practice and system-level variation affecting non-imaging data.

Third, technical tools for deidentification of medical images exist, and existing legal frameworks (for example, HIPAA in the USA) permit sharing such data. Different types of imaging present different challenges — the numeric time series that makes up ECG waveforms is trivial to deidentify, while a head magnetic resonance imaging (MRI) scan that could allow facial reconstruction is more complex — but these challenges are increasingly tractable (for example, all major cloud platforms offer a robust set of deidentification tools).

ground-truth patient outcomes that are necessary for researchers to solve non-trivial problems. Many health datasets available today implicitly treat human opinion as ground truth: an ECG is labeled with a cardiologist's judgment of arrhythmia, an X-ray is labeled with a radiologist's judgment on the severity of arthritis. While human labels are useful for efforts to automate human judgment, such efforts will also automate human biases and errors[12,13]. And ultimately, this approach is highly limiting: we want algorithms to do better than humans, not just produce the same results. To do so, we need algorithms that learn from nature — patient experiences and health outcomes — not physician judgment.

The task of creating ground-truth labels is not easy. Consider the task of labeling a biopsy image. It would be useful to know whether a patient ultimately progressed to metastatic cancer. But doing so, even when comprehensive electronic health records are available, requires a great deal of specialized knowledge: about cancer and where it metastasizes, how that event is recorded in the course of usual care, and how structural biases in health care affect when and how data are recorded. This places a major burden on individual researchers, particularly those without deep medical domain expertise. More problematic still, ground-truth labeling is also often infeasible in existing datasets: they can require dedicated efforts to link health system data to external sources of truth, for example, cancer registries or death records. Many health systems in the USA only record a patient's death if it happens within the four walls of the hospital — a problem given that only one-third of deaths in the USA occur in hospital. Linkages, for example to Social Security data in the USA or government registries elsewhere, can be essential but are neglected in many current datasets of health records.

Nightingale's accomplishments so far

By working closely with health systems, and investing in careful curation of data, Nightingale builds datasets that allow researchers to start asking and answering good questions quickly. At the time of launch, the platform housed five new imaging datasets (Box 1) totaling over 40 TB of images and waveforms, each focused on an important unsolved problem (Table 2).

All datasets are carefully documented not just with data dictionaries, but also a great deal of information on dataset construction and contents. This is responsive to growing awareness that some of the biggest problems in machine learning — failures to generalize, dataset shift, lack of representation and so on — happen when researchers overlook critical details of dataset creation.

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 a large urban hospital in Taiwan, and will soon expand to partnerships in Cameroon and Tamil Nadu.

How to interact with Nightingale

A key design principle of the platform is to minimize frictions in accessing the data, without compromising security or ethics. Practically, this means the process by which researchers gain access to the data is simple, easy and quick. No specific approval is required for research projects — a process that can take many months to complete in other health record datasets. Instead, users are approved on the basis of authenticating their identity, proving that they know how to handle potentially sensitive health data, and signing a data use agreement covering non-profit research (and non-profit research only). Following this, they are typically approved and can start working with data immediately. The interface is a familiar one for most researchers: a standard cloud computing environment with Jupyter notebooks and the ability to load nearly any package needed.

While the front-end experience is streamlined, the back end deploys a range of measures to ensure the highest security and ethical standards. First, Institutional Review Board-approved agreements cover all the research done to create and deidentify the datasets. Second, only deidentified data, certified by either our partner or by a third party under HIPAA Safe Harbor, are maintained on Nightingale. Third, and despite the lower risk of deidentified data, the data cannot leave: there is no download, and all access and analysis is done on the cloud computing platform Nightingale provides. Fourth, because the platform maintains total control of the data, every line of code executed on the platform can be surveilled and audited for compliance, giving immediate recourse against bad actors.

Legal and ethical barriers

We faced considerable skepticism about our ability to find hospital partners who would agree to any release of data, even deidentified. While many prospective partners declined to participate — citing privacy, insufficient funding, and a host of other reasons — we were encouraged to find a critical mass of institutions who shared our vision and signed on. Leaders and researchers at these systems were highly motivated by the value of these data to their patients and others around the world, as well as the prospect that the world's best computational researchers could be enticed to work on their problems, at no cost to them.

That said, we did face genuine challenges in our ability to successfully create open datasets as part of these collaborations. The first was a set of very understandable concerns regarding patient privacy. The deidentified nature of Nightingale datasets means that providing them to researchers is clearly allowed, under legal provisions for the sharing of such data for research in the USA and Europe (according to HIPAA and GDPR, respectively). Despite that, many internal legal teams adopted a far more restrictive interpretation than actually required by law.

Table 2 | Nightingale Open Science common task datasets

	Description	Common task example
Diagnosing 'silent' heart attack[14]	48,000 ECG waveforms linked to cardiac ultrasound reports	Diagnose regional wall motion abnormalities
Identifying high-risk breast cancer[15]	175,000 digital pathology slides linked to cancer registry data, treatments, and mortality from Social Security data	Predict breast cancer metastasis or death
Subtyping cardiac arrest[16]	24,000 ECG waveforms from cardiac arrest patients and matched controls, linked to post-arrest outcomes	Distinguish patients who go on to arrest versus normal controls
Predicting fractures[17]	64,000 chest X-rays linked to body measurements and diagnosed musculoskeletal conditions	Predict fracture risk
Emergency triage of patients with COVID-19 [18]	7,500 chest X-rays from patients with COVID-19, linked to in- and out-of-hospital measures of pulmonary deterioration	Predict intubation or death

A key lesson was to rely on templated institutional review board protocols and data use agreements, to which a range of other institutions had already agreed. This reassured internal legal teams, who were — very understandably — often reluctant to leap first, but content to do so if other peer institutions had already agreed. These templated agreements are publicly available on the Nightingale Open Science website. While useful, templates do not overcome a more fundamental problem: legal teams within health systems are often incentivized only to avoid downside risk to the institution, rather than to balance downside risk with benefits to patients — both those inside the health system, and in society more broadly. For this reason, it was very helpful to work directly with a researcher or leader in the health system who was empowered to balance these complex tradeoffs.

A broader source of resistance to data sharing came from the fact that most health systems lack a coherent framework that outlines the ethical — as opposed to the legal — basis for data sharing. In our discussions with health systems, we found it useful to invoke the principles from the Belmont Report as a guiding light for all such policy decisions. These principles, which are the foundation of the institutional review board review process, 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 foregrounding equity considerations.

Infrastructural barriers

A final, and perhaps surprising, challenge we faced relates to limitations of the infrastructure commonly used for storage and retrieval of high-dimensional data at health systems. On more than one occasion, we discovered that hospitals were bound by contracts with picture archiving and communication system (PACS) or ECG storage system vendors that imposed high per-image costs for export. These terms, which had not previously been appreciated before we started extracting data, made it prohibitively expensive for health systems to access the images, waveforms, and video generated in the course of patient care.

Another surprising discovery was that, in several important clinical settings, data are deleted or overwritten because of perceived storage space constraints. For example, in discussions with a top-ranked academic hospital, it became apparent that monitoring data from inpatient stays — ECG and pulse oximetry waveforms, high-frequency vital signs and so on, collected while patients occupied a monitored bed — were being overwritten with new data, starting 24 hours after the patient's discharge from the hospital. Similar practices were common in a wide range of settings, despite the negligible real costs of storing these data. While the deletion of data is anathema in many other industries, it is all too common in health care. Forward-looking health systems would do well to modify exploitative vendor contracts at the first opportunity, and ensure that no data are thrown away for arbitrary reasons.

What's next

There was a groundswell of interest from researchers in the months after launch, and Nightingale is already in use in classrooms at University of California, Berkeley, and Massachusetts Institute of Technology. To support our mission of the broadest possible access to data, Nightingale will soon be announcing a program to support researchers in under-resourced institutions and from under-represented groups in computer science, to cover computational costs and other expenses. It will also soon launch a grants program to solicit new datasets on important new medical common tasks, with a focus on problems and populations that are typically excluded from health datasets.

We hope creating open datasets anchored in common tasks, and providing them to the broadest possible community of researchers around the world, will give rise to a community of researchers who will form a new field: computational medicine. ❐

Sendhil Mullainathan[1] and
Ziad Obermeyer[2] ✉
[1]*Booth School of Business, University of Chicago, Chicago, IL, USA.* [2]*School of Public Health, University of California, Berkeley, Berkeley, CA, USA.*
✉*e-mail: zobermeyer@berkeley.edu*

Published online: 09 May 2022
https://doi.org/10.1038/s41591-022-01804-4

References

1. Chugh, S. S. *Circulation* **137**, 7–9 (2018).
2. Cancer data and statistics. *CDC* https://www.cdc.gov/cancer/dcpc/data/index.htm (2022).
3. COVID data tracker. *CDC* https://covid.cdc.gov/covid-data-tracker (2020).
4. Pierson, E. et al. *Nat. Med.* **27**, 136–140 (2021).
5. Beck, A. H. et al. *Sci. Transl Med.* **3**, 108ra113 (2011).
6. Ouyang, D. et al. *Nature* **580**, 252–256 (2020).
7. Yala, A. et al. *Sci. Transl Med.* **13**, eaba4373 (2021).
8. Donoho, D. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).
9. Kaushal, A. et al. *JAMA* **324**, 1212–1213 (2020).
10. McDermott, M. B. A. et al. *Sci. Transl Med.* **13**, eabb1655 (2021).
11. Price, W. N. & Cohen, I. G. *Nat. Med.* **25**, 37–43 (2019).
12. Mullainathan, S. & Obermeyer, Z. *AEA Papers & Proc.* **107**, 476–480 (2017).
13. Mullainathan, S. & Obermeyer, Z. *AEA Papers & Proc.* **111**, 37–42 (2021).
14. Pramanik, R. et al. Diagnosing 'silent' heart attack using ECG waveforms. *Nightingale Open Science Dataset* https://doi.org/10.48815/N54W2V (2021).
15. Bifulco, C. et al. Identifying high-risk breast cancer using digital pathology images. *Nightingale Open Science Dataset* https://doi.org/10.48815/N5159B (2021).
16. Huang, C. et al. Subtyping cardiac arrest with ECG waveforms. *Nightingale Open Science Dataset* https://doi.org/10.48815/N5WC7D (2021).
17. Lungren, M. et al. Predicting fractures and pain using chest x-rays. *Nightingale Open Science Dataset* https://doi.org/10.48815/N5RP44 (2021).
18. Robicsek, A. et al. Emergency triage of Covid-19 patients using chest X-rays. *Nightingale Open Science Dataset* https://doi.org/10.48815/N5MW26 (2021).

Acknowledgements

We thank K. Haynes, N. Foster, N. Juergens, W. Lane, K. Lin, J. Risley, and C. Simonian for their contributions to Nightingale Open Science. We are deeply grateful to T. Kalil for his vision and support, and thank T. Stewart for inspiring this Comment. We acknowledge financial support from Schmidt Futures, the Gordon and Betty Moore Foundation, Ken Griffin (private philanthropy), the National Institute on Aging grant P01AG005842, and the McGovern Foundation.

Author contributions

S.M. and Z.O. jointly wrote the paper.

Competing interests

S.M. and Z.O. have equity interests in LookDeep Health (healthcare services) and Dandelion (healthcare services).