PLOS DIGITAL HEALTH

REVIEW

Global healthcare fairness: We should be sharing more, not less, data

Kenneth P. Seastedt[1☯*], Patrick Schwab[2☯], Zach O'Brien[3☯], Edith Wakida[4☯], Karen Herrera[5☯], Portia Grace F. Marcelo[6☯], Louis Agha-Mir-Salim[7,8☯], Xavier Borrat Frigola[8,9☯], Emily Boardman Ndulue[10☯], Alvin Marcelo[11☯], Leo Anthony Celi[8,12,13☯]

1 Beth Israel Deaconess Medical Center, Department of Surgery, Harvard Medical School, Boston, Massachusetts, United States of America, **2** GlaxoSmithKline, Artificial Intelligence & Machine Learning, Zug, Switzerland, **3** Australian and New Zealand Intensive Care Research Centre (ANZIC-RC), Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Victoria, Australia, **4** Mbarara University of Science and Technology, Mbarara, Uganda, **5** Quality and Patient Safety, Hospital Militar, Managua, Nicaragua, **6** Department of Family & Community Medicine, University of the Philippines, Manila, Philippines, **7** Institute of Medical Informatics, Charité—Universitätsmedizin Berlin (corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health), Berlin, Germany, **8** Laboratory for Computational Physiology, Harvard-MIT Division of Health Sciences & Technology, Cambridge, Massachusetts, United States of America, **9** Anesthesiology and Critical Care Department, Hospital Clinic de Barcelona, Barcelona, Spain, **10** Department of Journalism, Northeastern University, Boston, Massachusetts, United States of America, **11** Department of Surgery, University of the Philippines, Manila, Philippines, **12** Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, United States of America, **13** Department of Biostatistics Harvard T.H, Chan School of Public Health, Boston, Massachusetts, United States of America

☯ These authors contributed equally to this work.
* kseasted@bidmc.harvard.edu

Citation: Seastedt KP, Schwab P, O'Brien Z, Wakida E, Herrera K, Marcelo PGF, et al. (2022) Global healthcare fairness: We should be sharing more, not less, data. PLOS Digit Health 1(10): e0000102. https://doi.org/10.1371/journal.pdig.0000102

Editor: Sanjay Aneja, Yale School of Medicine: Yale University School of Medicine, UNITED STATES

Published: October 6, 2022

Funding: LAC is funded by the National Institute of Health through NIBIB R01 EB017205. PS is an employee and shareholder of GlaxoSmithKline plc. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: Leo Anthony Celi is the Editor-in Chief of PLOS Digital Health.

Abstract

The availability of large, deidentified health datasets has enabled significant innovation in using machine learning (ML) to better understand patients and their diseases. However, questions remain regarding the true privacy of this data, patient control over their data, and how we regulate data sharing in a way that that does not encumber progress or further potentiate biases for underrepresented populations. After reviewing the literature on potential reidentifications of patients in publicly available datasets, we argue that the cost—measured in terms of access to future medical innovations and clinical software—of slowing ML progress is too great to limit sharing data through large publicly available databases for concerns of imperfect data anonymization. This cost is especially great for developing countries where the barriers preventing inclusion in such databases will continue to rise, further excluding these populations and increasing existing biases that favor high-income countries. Preventing artificial intelligence's progress towards precision medicine and sliding back to clinical practice dogma may pose a larger threat than concerns of *potential* patient reidentification within publicly available datasets. While the risk to patient privacy should be minimized, we believe this risk will never be zero, and society has to determine an acceptable risk threshold below which data sharing can occur—for the benefit of a global medical knowledge system.

Introduction

Many widely available imaging datasets exist containing deidentified data from thousands of patients and may be used to train machine learning (ML) algorithms, such as the COVID-19 Chest X-ray Dataset Initiative [1] and the CheXpert Chest Radiograph dataset [2]. In the ideal case, open datasets provide a robust and diverse foundation to train clinical prediction models, leading to improved predictive accuracy and generalizability of the derived models. However, questions remain regarding the true privacy of publicly available deidentified health data, patient control over their data, and how we regulate data sharing in a way that that does not encumber progress or further potentiate biases for underrepresented populations throughout the world.

In this review article, we address concerns over data anonymization and further explore how increased regulations may inadvertently exclude developing countries over concerns of imperfect data anonymization. We argue limiting data sharing would not only slow the development of future medical innovations and clinical software, but could also potentially expand existing biases that favor high-income countries. While the risks to patient privacy should be minimized, we believe this risk will never be zero, and an acceptable risk threshold below which data sharing can occur must be agreed upon by society for the benefit of a global medical knowledge system.

Deidentified health datasets promote innovation

The benefit of sharing deidentified data can be readily demonstrated by the widely used and publicly available Medical Information Mart for Intensive Care (MIMIC) database, now available in its fourth iteration [3,4]. This dataset includes deidentified clinical data from over 50,000 admissions to critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts, United States of America, spanning over a decade in time. Access to the database requires a "Data Use Agreement," which mandates that the developed source code for projects utilizing the data must be shared, promoting reproducibility and collaboration between research groups. Thousands of publications and conference proceedings have utilized this repository to advance our knowledge of critical care, and it has inspired the creation of similar databases in countries throughout the world, fostering international data sharing. The Society of Critical Care Medicine (SCCM) and the European Society of Intensive Care Medicine (ESICM) have embraced ICU patient data sharing [5], and the Amsterdam University Medical Center Database (AmsterdamUMCdb) adds to a growing list of globally available databases [6].

Further examples of how large international datasets have accelerated healthcare innovation continue to surface. One example includes a combination of mammography datasets from South Korea, USA, and the UK that led to the development of an AI algorithm that demonstrated not only improved breast cancer detection on mammography compared to radiologists, but also improved radiologist performance when assisted by AI [7]. Another example involves the use of open-access datasets and data from Stanford University—creating a dataset 2 orders of magnitude larger than previous skin pathology datasets—to develop a convolutional neural network (CNN) capable of skin cancer classification comparable to dermatology experts with potentially greater generalizability due to the size of the dataset enabling a more representative coverage of patients [8]. Similarly, combined datasets from China and the US were used to create algorithms capable of classifying macular degeneration and diabetic macular edema [9].

Shared learning from these large datasets continue to provide powerful potential to enable scientific advances and medical innovation. However, many barriers exist to the creation of

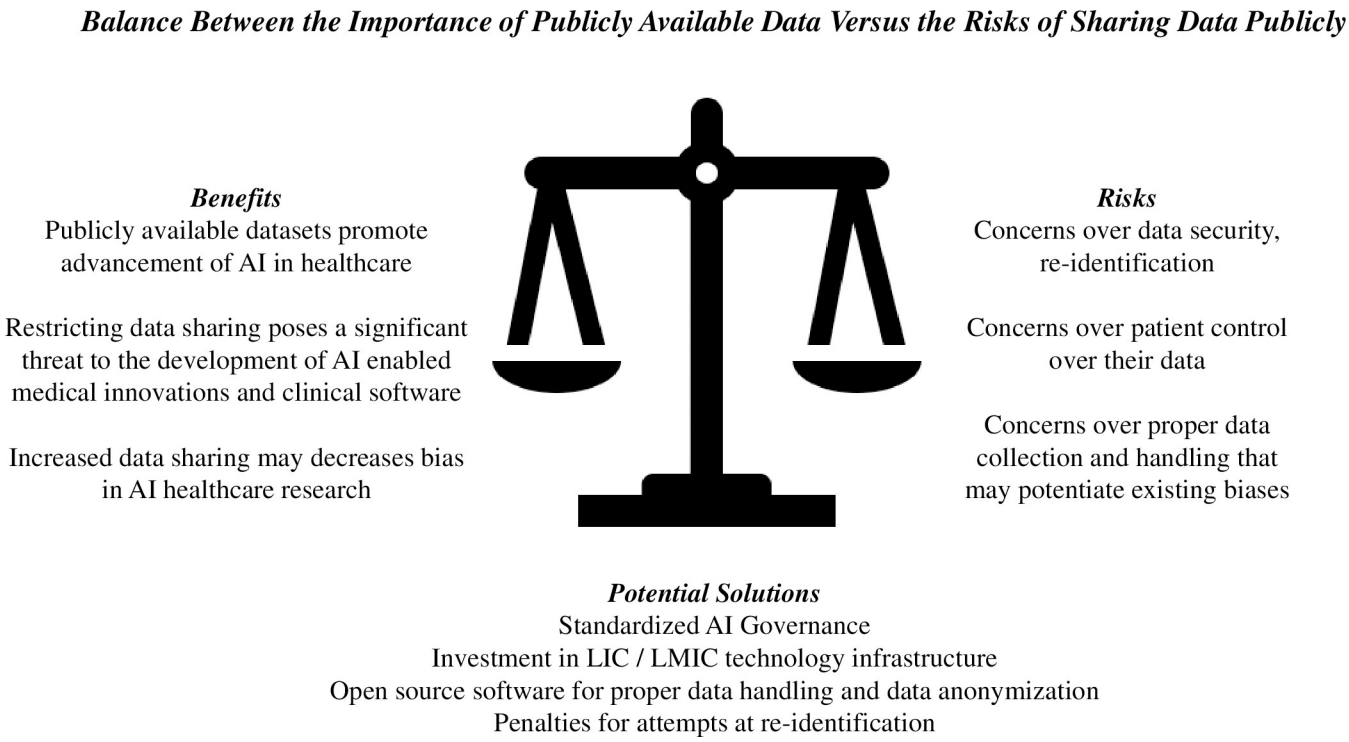

Fig 1. The balance between the importance of publicly available data versus the risks of sharing data publicly. Potential solutions are presented. AI, artificial intelligence; LIC, low-income country; LMIC, low-middle-income country.

https://doi.org/10.1371/journal.pdig.0000102.g001

large, publicly available datasets, including concerns over the security of the shared data and how the open dataset will be used (Fig 1).

Concerns over publicly available dataset security and regulation

In the ideal case, publicly available, deidentified datasets provide a robust and diverse foundation to train clinical prediction models, improving predictive accuracy and generalizability of the derived models and enabling medical innovation. However, concerns over the security and privacy of these datasets exist, potentially limiting the creation and diversity of new datasets. Interest in governmental regulation of these datasets is becoming more prevalent, as demonstrated by proposed laws in the US [10], China [11], and the European Union [12]. Patients are also becoming increasingly aware that their data are being used for research or commercial purposes and are rightfully more interested (particularly women and minorities) in controlling that data and how it is used [13]. One study found that most patients want to know what their personal health information (PHI) will be used for and are uncomfortable sharing their PHI with commercial entities but are comfortable sharing with their own institution [14]. Younger patients and women preferred more control of their PHI than older patients and males with respect to research participation, with a significant number of patients preferring study-specific consents. Most wanted to be informed of what their PHI was being used for, and 70% wanted to receive the results of studies using their PHI—highlighting the shifting patient preferences for more transparency and interest in how their PHI is being used.

Concerning data security, best practices [15] exist in the US for deidentifying image data to comply with the standards outlined in the Health Information Portability and Accountability Act (HIPAA) Privacy Rule [16], protecting a patient from potentially being identified from

publicly available images. As defined in the HIPAA Privacy Rule, deidentified data are not regulated and may be shared without restriction. Standard deidentification techniques to remove protected information from images and associated metadata include pseudo-anonymization and k-anonymity [17,18]. However, deidentification methods are often automated and can be imperfect, as demonstrated, for example, by Google canceling the release of a public chest X-ray (CXR) dataset after discovering patient data was still embedded in some of the images [19].

Beyond the risk of leaking private information directly with an insufficiently deidentified dataset, there also exists the risk that an attacker may attempt to utilize information present in other publicly or privately available datasets to reestablish a link between deidentified patient images and their individual identities. One group has recently presented an ML approach that further highlights the linkage risk inherent in publicly available imaging datasets. Instead of training models on the available data to potentially predict pathology, they trained deep-learning (DL) models to identify patients from the available images. Their results are revealing: Multiple images of a single patient *can* accurately be determined to belong to the same original patient, despite standard deidentification efforts and without a shared identifier linking these images together [20]. Using 2 Siamese neural networks (SNN) trained on the Chest X-ray 14 dataset [21], the authors were able to determine whether 2 CXRs belonged to the same patient, even if taken several years apart and with new pathological development between the time points the images were taken. The implications are far-reaching: Given a patient's CXR image, one could potentially match that image to other publicly or privately available CXR datasets that may contain imperfectly deidentified metadata and reidentify that patient or gain access to additional sensitive information. The ability to accurately match patients across deidentified datasets exposes the weakness of relying on deidentification techniques that do not guarantee complete anonymization or differential privacy [22]. Beyond reidentification from CXR images, additional concerns exist regarding reidentification from head and neck images that include the patient's face, highlighting the need for defacing software to deface patient images in datasets that include facial profiles [23,24].

An important consideration is the trade-off between the degree of anonymization—as measured in terms of differential privacy—and utility/representativeness for downstream clinical prediction tasks [25]. Despite Packhauser and colleagues demonstrating a potential route for an attacker to gain access to sensitive patient information, it is essential to note that the mere ability to match records belonging to the same patient does not yet constitute a reidentification. An attacker would still require access to either (i) an imperfectly deidentified dataset that allows further inferences about the patient or their identity or (ii) a dataset that was not deidentified containing private information about the patient to be able to exploit the ability to match patients across datasets [26].

Tempering concerns over public data security

Despite these valid concerns over data security, there currently exists little publicly available evidence of patient identities having been linked to open health data (OHD) despite the theoretical impossibility of true anonymization. "Nothing about an individual should be learnable from the database that cannot be learned without access to the database" was, to the best of our knowledge, first proven by Dwork [27] and has led to the development of the differential privacy framework that—in lieu of full anonymization—instead seeks to quantify and provide theoretical limits for the maximum privacy loss incurred by individuals. Dwork's impossibility result has far reaching consequences for healthcare practitioners wishing to release *any* data. In particular, reidentification attacks, such as the one highlighted in [20], cannot be ruled out fully due to auxiliary information, and practitioners wishing to release healthcare data are

therefore left with the difficult decision of managing the trade-off between the value to society of sharing data on the one hand and the risk of privacy loss for individuals on the other. While the previously mentioned HIPAA Privacy Rule outlines clear criteria for deidentification, it does not provide regulatory guidelines on managing differential privacy trade-offs. It is important to note that the potential for privacy loss and reidentification applies to the release of any data, including statistics on a cohort level. In response to these challenges, several organizations that routinely deal with personal data and statistics thereof have turned towards the adoption of differential privacy methods to systematically quantify and manage privacy risks, such as, for example, the US Census Bureau [28] and Apple [29].

To better quantify the risk of patient reidentification, we sought to evaluate the literature on this topic to assess current concerns over data security, including larger-scale data breaches. Subtle approaches to reidentification of (potentially improperly) anonymized health data stand in stark contrast to the illegal, forcible acquisition of personal health data by means of a data breach—which includes illegal disclosure, attainment, or use of information without authorization. Theft of medical records—in contrast to credit card records—is attractive to criminals because they contain sufficient information to secure loans, open a bank account, obtain health services and prescription medication, etc. In brief, identity theft may be the principal reason for intentional data disclosure [30]. Of reported data breaches in the USA between 2015 and 2019, the health sector accounted for 76% of all 10 billion cases [31].

Under US federal legislation, if a healthcare data breach affects 500 or more patients, it must be reported to the Office of Civil Rights (OCR). The OCR data breach portal provides an online database describing data breaches of protected health information (PHI) [32]. To evaluate the frequency of PHI data leakage through data breaches, we downloaded and analyzed data containing type of breach, location of breached information, and number of individuals affected during the last 2 years. The most frequent type of PHI data breach was the one produced by hacking activities (72%) followed by unauthorized disclosure (21%). Regarding where the data was hosted when the breach occurred, almost all leaks affected servers (93%) and the second location was email (35%). Network servers and email have become the main locations for hackers using different techniques such as malware, ransomware, or phishing attacks to prey on electronic health records (EHRs) [31]. Note that 62% of the data leaks came from more than 1 location. Combining type of breach and location, hacking the internal network to reach servers is the main cause of PHI data leakage well above other causes (Fig 2).

Compared to the large and growing problem of hackers attacking PHI data servers, there is no category nor series of case reports about data leakage from OHD. A potential reason for this discrepancy could be that anonymized OHD would generally be stripped of patients' full names, address histories, financial information, and social security numbers, making OHD less valuable for criminal purposes than private datasets that include personalized information. Solely clinical data and "quasi-identifiers" remain that could potentially be reidentified with relatively greater effort when combined with other data sources that contain personal identifiers. In the end, a threat by criminal activity is present wherever data are stored and are always subject to misuse or theft—with the utility of the data for malicious uses being diminished the harder establishing a link to personalized identifiers is.

In order to approximate the number of known cases where individuals were reidentified from OHD, we performed a literature review on PubMed [query = ("case" OR "example") AND ("re-identification" OR "reidentification" OR "re-id") NOT ("theoretical" OR "video" OR "camera" OR "pedestrian" OR "visual")], excluding studies on reidentification attempts from security cameras and genomic information (S1 Table). Reviewing the literature (n = 65) and other relevant articles, reports on individuals being publicly reidentified from OHD were scarce. A systematic review of reidentification attacks concluded the success rate of health data

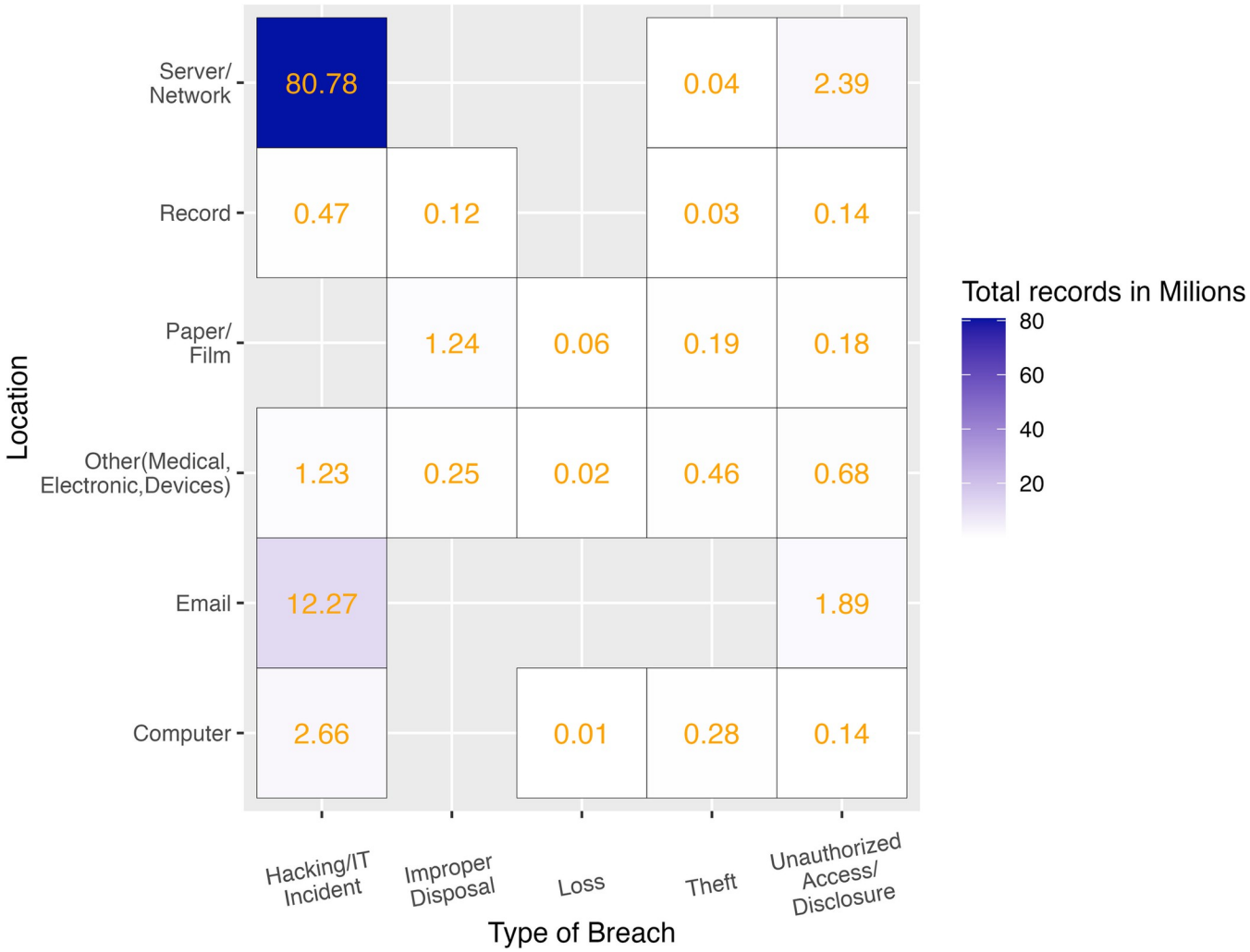

Fig 2. Magnitude of data breaches considering total number of records affected by type of breach and location in the US between November 2019 and November 2021. Elaborated from raw data [32].

https://doi.org/10.1371/journal.pdig.0000102.g002

reidentification attempts was very low when considering studies that included data deidentified according to modern standards [33]. A contemporary analysis that assessed the reidentification risk from a healthcare dataset of over 1 million patients at Vanderbilt University Medical Center also suggests that reidentification risk is low considering a potential attacker's resources and capabilities [34].

In order to get a more holistic view outside of the medical literature, we expanded our search by reviewing news coverage using Media Cloud, an open-source global news database and analysis tool [35]. Media Cloud covers news stories, blog entries, and web pages and regularly ingests news content from more than 60,000 publications worldwide. It also provides retrospective coverage for many sources going back to 2010. Our search was limited to the US only, as the use of this system requires careful selection of a media corpus to study. The well-bounded media "collection" of the US provides the ability to draw meaningful conclusions about the volume and proportion of attention to a specific topic within a larger media coverage "universe" instead of as a tool to simply find relevant articles. Including other global areas would have potentially diluted our ability to find meaningful results using this method, and

the US media ecosystem is also known to be sensitive for media attention to subjects globally. For this survey, we searched over 10,000 media publications from the US, publishing at either the state or national level. The query used to identify coverage on this issue was ["de-anonymization" OR deanonymization OR "de-anonymize" OR deanonymize OR "re-identification" OR reidentification OR "re-identify" OR reidentify]. The timeframe of coverage searched was 5 years, 09-01-2016 through 09-01-2021. Data were manually cleaned to remove a minimal number of irrelevant articles (in which the term deanonymize was used in other contexts), or coverage to the issue outside of the US. The resulting corpus was 186 stories from 127 sources (S2 Table). We manually coded the articles on 2 variables: The first was the context in which the issue was discussed: theoretical discussion (143/186, 77%), discussion of research released (20/186, 11%), or discussion of an actual case (23/186, 12%). The remaining articles without cases or research were assigned the code of discussing the issue theoretically (i.e., in the abstract). Despite ongoing thematic coverage of this issue over the past 5 years, there is a paucity of actual examples (S3 Table). Hence, we can assume that if there were more publicly known examples, there would be considerable coverage given the ongoing news attention to this issue. None of the identified cases involved health data nor healthcare databases.

Against this backdrop, the medical field must address essential questions, such as how do we improve PHI protection, how do we increase patient involvement in how their data are used, and how do we do this in a way that continues to promote global collaborative efforts in analyzing OHD without a complete shutdown in progress? Given the paucity of evidence supporting concerns over data security of publicly available datasets, we believe that continued investment in publicly available datasets to promote innovation is prudent and that there may be implicit harm in limiting data sharing.

Potential harms of prematurely limiting data sharing—Potentiating bias

The aforementioned developments suggest that matching deidentified patient records across datasets is potentially easier than previously thought, and patients are at the same time increasingly interested in more control over how their data are being used. Increased availability of deidentified patient data has led to a global boom [36] in innovation with ML-driven software as a medical device (SaMD), with little data to support concerns over publicly available medical data security. Crucial to this discussion is how limiting data sharing (such as the current legal framework proposed by the EU [12]) would affect underrepresented populations and potentiate bias [37].

The medical knowledge system that informs clinical practice worldwide has historically been based on studies primarily performed on a handful of high-income countries and typically enrolling white males [38–40]. Guidelines for the management of heart disease, for example, are disseminated to the rest of the world from professional societies such as the American Heart Association. To truly move towards a global knowledge medical system that incorporates data from all parts of the world to decrease bias and increase data fairness, data from places that historically have not had a leading role in the development of current medical standards should be included—ethnic minorities, lower-income countries (LICs), and lower-middle-income countries (LMICs). One example in which bias affected an ML algorithm includes a breast cancer histology algorithm that reflected ethnicity rather than intrinsic tumor biology due to site-specific staining protocols and region-specific demographics [41]. In this example, the bias was introduced due to 1 site having more black patients included than many of the other sites. Biased models risk repeating cancer care inequities related to ethnic background. Another study identified extensive bias in several publicly available CXR datasets used for ML

algorithms and found multisource datasets may combat such bias [42]. There is evidence of bias and underrepresentation even within the US as 70% of data comes from 3 wealthy states as opposed to rural regions, and less than one-third of states are represented in data-sharing platforms [43].

With digitalization, every country has an opportunity to create its own medical knowledge system using data routinely collected in the process of care. However, many countries are only just starting to reach the levels of digitalization of countries like the US and China, and increased regulations on the use of AI could further limit the participation of developing countries in these global datasets. Up to 94% of funding for AI startups over the last 5 years is accounted for by the US and China [44]. This poses the risk of potentiating bias given the limited diversity of datasets. However, there could potentially be harm in using data from developing countries as well given concerns for poor data collection methods and lack of inclusion of disadvantaged populations. Indeed, biased AI has led to racial profiling in South Africa [45] and labor exploitation in Venezuela [46]. In Africa and Latin America, there is a significant lack of knowledge and regulatory frameworks surrounding PHI, and the concept of PHI is largely alien to the patient and sometimes practitioners [47,48]. Similarly, in the Philippines, illness and care is a communal experience, with many taking comfort in sharing the steps of their care process with family and close friends—which, in many rural areas without digital healthcare, is essentially the entire village or *barangay*. As many as 2.3 million Filipino families have no electricity, limiting access to digital healthcare. Sub-Saharan Africa (including Uganda) suffers from limited usage of electronic health records due to the high cost of procurement and maintenance, poor internet connectivity, intermittent power supply to the rural settings, and low uptake by healthcare workers [49,50]. There is limited training on how to protect and/or handle patient information as this is prioritized secondary to care delivery given technical challenges (e.g., power supply, internet, and computer infrastructure) and the requirement that a strained workforce must serve high volumes of patients [51].

The digital healthcare experience in these countries emphasizes how increased regulations on publicly available datasets will likely raise the barrier to entry for developing countries, further excluding their populations from datasets and increasing biases that favor high-income countries. While no data directly supports this at this time, there is some evidence that limiting data flow adversely affects innovation [52]. It is possible that barriers to data flow make it more time-intensive and expensive to share data overseas, benefiting those countries with the resources to overcome these barriers.

Potential solutions

Therefore, if we are to achieve unbiased datasets that represent the global community, the leaders in healthcare digitization need to assist LICs/LMICs with contributing to publicly available datasets, but also assist in enabling accurate data collection. This support would allow these countries to leverage their data to solve clinical problems unique to their populations and improve current global datasets by more representatively covering diverse populations. As LICs and LMICs embark on developing digital health infrastructure, views of the marginalized and vulnerable must be included in defining how their data will be collected, used, and how they can benefit. It is therefore essential to engage patients and the community when promoting digital literacy in LICs/LMICs. Varied demographics—for example, laborers, the urban poor, indigenous peoples, elderly, women—must be involved in developing PHI governance bodies so that their voices are included.

Other proposed solutions to improving data sharing include promising new technologies, such as synthetic data or federated learning [53–55], which have been suggested to potentially

help researchers publicly sharing health data while better managing the risk of deidentification. However, linkage risk will always remain a concern as even releasing summary statistics alone constitutes a certain loss of privacy for the contributing data sources in terms of differential privacy [28,56]. In the federated learning framework, investigators from different institutions combine efforts by training a model locally on their own data, and sharing the trained model parameters with others to generate a central model, rather than sharing the source data directly [57]. However, the feasibility of using federated learning for data sharing is predicated on consistent data curation, standards, and harmonization across the participating institutions. Additionally, given that the data are not combined, the opportunity to expand the number of rare events may not be fully leveraged if the modeling is performed in isolation and only the meta-model is shared across institutions. Most importantly, algorithmic bias will be harder to detect if local investigators only see their own data. Given how challenging it is to detect and fix algorithmic bias in models trained on pooled data, it would likely be even more difficult to address algorithmic bias when learning is distributed. Resource allocation towards federating learning platforms and technologies should therefore be balanced with those allocated towards better tools for deidentification and standardized data curation.

To ensure proper data collection and sharing, legal policy and data security frameworks should be put in place to strengthen the protection of PHI datasets from accidental leakage and potential malicious outside attacks [58]. These policies should in particular regulate stewards of PHI datasets that, if combined, may enable reidentification via linkage with openly available datasets. Substantial penalties should be developed to punish any attempts to exploit linkage of open medical data with the aim of reidentifying patients or using PHI for commercial purposes, rather than for society's benefit, without patient consent. Additionally, although increasing patient involvement undoubtedly adds more complexity, patient stewardship over their data is a fundamental right. As the technology to study PHI advances, technologies to improve PHI management ought to advance in lockstep. Numerous countries have embarked on creating AI governance frameworks, but there is no central coordination between nations to set standards for proper handling of data sharing across international boundaries [59]. Equitable AI governance and attempts at global AI regulations and standards may help consider the needs and inequalities of developing countries [60].

Investment in technology infrastructure (such as EHRs) for data collection and data sharing and surveillance for this technology should be a priority for these populations. The benefits of global investment in LIC/LMIC digitalization are numerous and include improved accuracy of collected data through health information management systems, decreased bias, and improved algorithmic fairness through the inclusion of marginalized groups in training data and ensure accountability for proper data collection and sharing. To foster investment, developing countries may want to consider incentives for foreign tech companies to conduct research and develop facilities to promote infrastructure development. The global AI community should continue to consider these investments, creating open-source software to promote proper data handling, data anonymization software, and AI governance standards. Although MIMIC represents a model for the use of big data and how it may contribute to improving medical understanding in high-income countries, such a model may be less feasible in LICs/LMICs where patients may lack access to advanced clinical services. Thus, as health systems in LIC/LMICs undergo digital transformation, there should be equal attention to, or even affirmative action towards, data analysis of services rendered at the primary care level where the majority of clinical encounters for health promotion and disease prevention occur. In many cases, poor or socially disadvantaged patients may have more complex diseases and have no recourse but to receive care in the nearest public primary care facility, potentially never reaching a hospital. These patients would not be represented if data collection was limited to

hospitalized patients, and, therefore, the structure of local healthcare systems must be considered when designing open clinical databases for LICs/LMICs.

Conclusions

We would argue that the cost—measured in terms of access to future medical innovations and clinical software while potentiating bias—of slowing ML progress is too great to stop sharing data through large publicly available databases for concerns over imperfect anonymization and potential linkage risks. Although the potential for linking public medical records at the detriment of patients exists, a robust regulatory framework that protects both open sharing of deidentified data for good, and strongly penalizes patient reidentification, may be a more measured solution than attempting to broadly limit data sharing. Publicly available datasets provide the fuel for widespread application and adoption of AI in healthcare and for advancing our understanding of heterogeneous and diverse patient populations globally. Slowing progress by limiting data sharing risks curtailing medical innovation and significantly impeding our ability to advance our understanding of health and global disease. Preventing AI's progress towards precision medicine and sliding back to the "white-size-fits-all" clinical practice dogma poses a more significant threat than contemporary concerns of *potential* patient reidentification within publicly available datasets. This potential reidentification risk will never be zero, and we have to determine an acceptable risk threshold for sharing data for the benefit of a more global medical knowledge system. The global AI community needs to take an active role to assist developing nations on their healthcare digitization quest through standardized AI governance for data sharing and investment in equitable data collection infrastructure.

Supporting information

S1 Table. Open health data reidentification PubMed review results.
(PDF)

S2 Table. Coded deanonymization and reidentification stories.
(PDF)

S3 Table. Complete list of individual cases on media about personal information disclosure.
(PDF)

References

1. COVID-19 Chest X-Ray Dataset Initiative. Available from: https://github.com/agchung/Figure1-COVID-chestxray-dataset. [cited Mar 2021].
2. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al., editors. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence; 2019.
3. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 1.0). PhysioNet. 2021. https://doi.org/10.13026/s6n6-xd98.
4. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016; 3(1):160035. https://doi.org/10.1038/sdata.2016.35 PMID: 27219127
5. Kaplan LJ, Cecconi M, Bailey H, Kesecioglu J. Imagine. . .(a common language for ICU data inquiry and analysis). Intensive Care Med. 2020; 46(3):531–3. Epub 2020/03/04. https://doi.org/10.1007/s00134-019-05895-5 PMID: 32123991.
6. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database

(AmsterdamUMCdb) Example. Crit Care Med. 2021. Epub 2021/02/25. https://doi.org/10.1097/ccm.0000000000004916 PMID: 33625129.

7. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health. 2020; 2(3):e138–e48. Epub 20200206. https://doi.org/10.1016/S2589-7500(20)30003-0 PMID: 33334578.
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639):115–8. Epub 20170125. https://doi.org/10.1038/nature21056 PMID: 28117445; PubMed Central PMCID: PMC8382232.
9. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018; 172 (5):1122–31.e9. https://doi.org/10.1016/j.cell.2018.02.010 PMID: 29474911
10. Johnson E. Text—H.R.6216 - 116th Congress (2019–2020): National Artificial Intelligence Initiative Act of 2020. (2020 March 12). Available from: http://www.congress.gov/.
11. State Council. Notice of the State Council Issuing the New Generation of Artificial Intelligence Development Plan. State Council Document [2017] No. 35, 2017. Available from: https://www.unodc.org/res/ji/import/policy_papers/china_ai_strategy/china_ai_strategy.pdf.
12. European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Brussels, 4/21/21. Available from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206.
13. Willison DJ, Swinton M, Schwartz L, Abelson J, Charles C, Northrup D, et al. Alternatives to project-specific consent for access to personal information for health research: Insights from a public dialogue. BMC Medical Ethics. 2008; 9(1):18. https://doi.org/10.1186/1472-6939-9-18 PMID: 19019239
14. Tosoni S, Voruganti I, Lajkosz K, Habal F, Murphy P, Wong RKS, et al. The use of personal health information outside the circle of care: consent preferences of patients from an academic health care institution. BMC Medical Ethics. 2021;22(1). https://doi.org/10.1186/s12910-021-00598-3 PMID: 33761938
15. Moore SM, Maffitt DR, Smith KE, Kirby JS, Clark KW, Freymann JB, et al. De-identification of medical images with retention of scientific research value. Radiographics. 2015; 35(3):727–735. https://doi.org/10.1148/rg.2015140244 PMID: 25969931
16. Centers for Disease Control and Prevention (CDC). HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. MMWR: Morbidity and mortality weekly report. 2003; 52(Suppl 1):1–17, 9.
17. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing Medical Imaging Data for Machine Learning. Radiology. 2020; 295(1):4–15. https://doi.org/10.1148/radiol.2020192224 PMID: 32068507
18. Aryanto KYE, Oudkerk M, Van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol. 2015; 25(12):3685–3695. https://doi.org/10.1007/s00330-015-3794-0 PMID: 26037716
19. MacMillan D, Bensinger G. Google almost made 100,000 chest X-rays public—until it REALIZED personal data could be exposed. 2019, November 18. Available from: https://www.washingtonpost.com/technology/2019/11/15/google-almost-made-chest-x-rays-public-until-it-realized-personal-data-could-be-exposed/. [cited Mar 2021].
20. Packhauser K, Gundel S, Munster N, Syben C, Christlein V, Maier A. Is Medical Chest X-ray Data Anonymous? arXiv pre-print server. 2021. https://doi.org/10.48550/arXiv.2103.08562
21. Wang X et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classificationand Localization of Common Thorax Diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017;2097–2106.
22. Dwork C, editor Differential privacy: A survey of results. International conference on theory and applications of models of computation; 2008: Springer.
23. Crawford KL, Neu SC, Toga AW. The Image and Data Archive at the Laboratory of Neuro Imaging. NeuroImage. 2016; 124:1080–1083. https://doi.org/10.1016/j.neuroimage.2015.04.067 PMID: 25982516
24. Image defacing using BioImage Suite Web. Available from: https://bioimagesuiteweb.github.io/bisweb-manual/tools/defacing.html. [cited 2022 Jun 12].
25. Cheng V, Suriyakumar VM, Dullerud N, Joshi S, Ghassemi M. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.
26. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010; 17 (2):169–177. https://doi.org/10.1136/jamia.2009.000026 PMID: 20190059

27. Dwork C, editor. Differential Privacy 2006; Berlin, Heidelberg: Springer Berlin Heidelberg.
28. Abowd JM. The U.S. Census Bureau Adopts Differential Privacy2018. ACM.
29. Tang J, Korolova A, Bai X, Wang X, Wang X. Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12. arXiv pre-print server. 2017. https://doi.org/10.48550/arXiv.1709.02753
30. Coventry L, Branley D. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas. 2018; 113:48–52. https://doi.org/10.1016/j.maturitas.2018.04.008 PMID: 29903648
31. Seh AH, et al. Healthcare Data Breaches: Insights and Implications. Healthcare (Basel, Switzerland) 8, (2020). https://doi.org/10.3390/healthcare8020133 PMID: 32414183
32. Office for Civil Rights U.S. Department of Health & Human Services. Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information. Available from: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.
33. Emam KE, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS ONE. 2011; 6 (12):e28071. https://doi.org/10.1371/journal.pone.0028071 PMID: 22164229
34. Xia W, Liu Y, Wan Z, Vorobeychik Y, Kantacioglu M, Nyemba S, et al. Enabling realistic health data re-identification risk assessment through adversarial modeling. J Am Med Inform Assoc. 2021; 28 (4):744–752. https://doi.org/10.1093/jamia/ocaa327 PMID: 33448306
35. Roberts H, Bhargava R, Valiukas L, et al. Media cloud: massive open source collection of global news on the open web. Proceedings of the International AAAI Conference on Web and Social Media. 2021;15:1034–1045.
36. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020; 3(1):1–8.
37. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. Lancet Glob Health. 2021; 3(4):e260–e265. https://doi.org/10.1016/S2589-7500(20)30317-4 PMID: 33678589
38. Czerniewicz L. It's time to redraw the world's very unequal knowledge map. Available from: https://theconversation.com/its-time-to-redraw-the-worlds-very-unequal-knowledge-map-44206. [cited April 2021]. 2015.
39. Sacasas F, A J. Controversias en torno a la medicina basada en evidencias. Revista Habanera de Ciencias Médicas. 2011; 10 (3):339–347.
40. Niranjan SJ, Durant RW, Wenzel JA, Cook ED, Fouad MN, Vickers SM, et al. Training Needs of Clinical and Research Professionals to Optimize Minority Recruitment and Retention in Cancer Clinical Trials. J Cancer Educ. 2019; 34(1):26–34. https://doi.org/10.1007/s13187-017-1261-0 PMID: 28776305
41. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021; 12(1):4423. Epub 20210720. https://doi.org/10.1038/s41467-021-24698-1 PMID: 34285218; PubMed Central PMCID: PMC8292530.
42. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. arXiv:200300827 [cs, eess, stat] [Internet]. 2020 Oct 15 [cited 2022 Jun 12]; Available from: http://arxiv.org/abs/2003.00827.
43. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA. 2020; 324:1212–1213. https://doi.org/10.1001/jama.2020.12067 PMID: 32960230
44. United Nations Conference on Trade and Development Digital economy report 2021. "Cross border data flows and development: for whom the data flow." 6/12/21. Available from: https://unctad.org/system/files/official-document/der2021_en.pdf
45. Hao K, Swart H. South Africa's private surveillance machine is fueling a digital apartheid. MIT Technology Review Available from: https://www.technologyreview.com/2022/04/19/1049996/south-africa-ai-surveillance-digital-apartheid/.
46. Hao K, Hernandez A. How the AI industry profits from catastrophe [Internet]. MIT Technol Rev. 2022. Available from: https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/.
47. Olivier MS. Database privacy: balancing confidentiality, integrity and availability. SIGKDD Explor Newsl. 2002; 4(2):20–27.
48. Solove DJ. Conceptualizing privacy. Calif L Rev. 2002; 90:1087.
49. Akanbi MO et al, Use of Electronic Health Records in sub-Saharan Africa: Progress and challenges. J Med Trop. 2012; 14(1):1–6. PMID: 25243111; PMCID: PMC4167769.
50. Izaara AA, Ssembatya R, Kaggwa F. An access control framework for protecting personal electronic health records. In 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC) (pp. 1–6). IEEE.). 2018, December.

51. Kiberu VM, Matovu JK, Makumbi F, Kyozira C, Mukooyo E, Wanyenze RK. Strengthening district-based health reporting through the district health management information software system: the Ugandan experience. BMC Med Inform Decis Mak. 2014; 14(1):1–9. https://doi.org/10.1186/1472-6947-14-40 PMID: 24886567

52. Cory N, Dascoli L. How barriers to cross-border data flows are spreading globally, what they cost, and how to address them. 2021. Available from: https://itif.org/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost/null/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost/

53. Jordon J, Yoon J, Van Der Schaar M, editors. PATE-GAN: Generating synthetic data with differential privacy guarantees. International Conference on Learning Representations; 2018.

54. Schütte AD, Hetzel J, Gatidis S, Hepp T, Dietz B, Bauer S, et al. Overcoming Barriers to Data Sharing with Medical Image Generation: A Comprehensive Evaluation. arXiv preprint arXiv:201203769. 2020.

55. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Rader B, Ingerman A, et al. Privacy-first health research with federated learning. 2020.

56. Dwork C. (2006) Differential Privacy. In: Bugliesi M., Preneel B., Sassone V., Wegener I. (eds) Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer, Berlin, Heidelberg https://doi.org/10.1007/11787006_1

57. Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. A systematic review of federated learning applications for biomedical data. PLoS Digit Health. 2022; 1(5):e0000033. https://doi.org/10.1371/journal.pdig.0000033

58. Teague V. The Simple Process of Re-Identifying Patients in Public Health Records. Available from: https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records. [cited Jan 2022].

59. Radu R. Steering the governance of artificial intelligence: national strategies in perspective. Polic Soc. 2021; 40 (2):178–193. https://doi.org/10.1080/14494035.2021.1929728

60. Paris Peace Forum. Beyond the North-South Fork on the Road to AI-Governance: An Action Plan for Democratic & Distributive Integrity. 2022. Available from: https://digitalrights.ai/report/.