




# Delivering on NIH data sharing requirements: avoiding Open Data in Appearance Only

Hope Watson,<sup>1</sup> Jack Gallifant ,<sup>2</sup> Yuan Lai,<sup>3,4</sup> Alexander P Radunsky,<sup>5,6,7</sup> Cleve Villanueva,<sup>8</sup> Nicole Martinez,<sup>9</sup> Judy Gichoya ,<sup>10</sup> Uyen Kim Huynh,<sup>11</sup> Leo Anthony Celi <sup>12,13</sup>

**To cite:** Watson H, Gallifant J, Lai Y, *et al.* Delivering on NIH data sharing requirements: avoiding Open Data in Appearance Only. *BMJ Health Care Inform* 2023;**30**:e100771. doi:10.1136/bmjhci-2023-100771

Received 23 March 2023  
Accepted 29 May 2023

## ABSTRACT

**Introduction** In January, the National Institutes of Health (NIH) implemented a Data Management and Sharing Policy aiming to leverage data collected during NIH-funded research. The COVID-19 pandemic illustrated that this practice is equally vital for augmenting patient research. In addition, data sharing acts as a necessary safeguard against the introduction of analytical biases. While the pandemic provided an opportunity to curtail critical research issues such as reproducibility and validity through data sharing, this did not materialise in practice and became an example of 'Open Data in Appearance Only' (ODIAO). Here, we define ODIAO as the intent of data sharing without the occurrence of actual data sharing (eg, material or digital data transfers).

**Objective** Propose a framework that states the main risks associated with data sharing, systematically present risk mitigation strategies and provide examples through a healthcare lens.

**Methods** This framework was informed by critical aspects of both the Open Data Institute and the NIH's 2023 Data Management and Sharing Policy plan guidelines.

**Results** Through our examination of legal, technical, reputational and commercial categories, we find barriers to data sharing ranging from misinterpretation of General Data Privacy Rule to lack of technical personnel able to execute large data transfers. From this, we deduce that at numerous touchpoints, data sharing is presently too disincentivised to become the norm.

**Conclusion** In order to move towards Open Data, we propose the creation of mechanisms for incentivisation, beginning with recentring data sharing on patient benefits, additional clauses in grant requirements and committees to encourage adherence to data reporting practices.

## INTRODUCTION

Six years on from the development of the FAIR data principles<sup>1</sup> (Findability, Accessibility, Interoperability, and Reusability), the recent deployment of the NIH data sharing mandate is a significant step towards increasing the reproducibility and robustness of research that has long eluded the data science community.<sup>2-4</sup> From January 2023, NIH intramural investigators will be required to prospectively

plan for the management and sharing of scientific data, and must submit a data management and sharing (DMS) for each new grant.<sup>2</sup> At a minimum data supporting a publication must be shared at the time of dissemination, and other scientific data released at the end of the research project or protocol. This mandate facilitates an ecosystem-wide shift in mindset surrounding data sharing, creating a culture that places efficient accumulation of knowledge and, ultimately, patients first.<sup>5,6</sup>

Unfortunately, previous initiatives have encountered several obstacles and resistance, as data sharing is not as simple as is often implied.<sup>7</sup> The COVID-19 pandemic highlighted this issue, demanding the public reporting of health data at a scale unlike any other. Continuous monitoring of the quality of care and international comparisons was vital.<sup>8</sup> The clinical and academic communities were also desperate for patient-level data that researchers could evaluate to identify trends and treatments. A significant volume of preprints of questionable reliability transpired, which had no way of validating results.<sup>9</sup> Furthermore, there has been a lack of precise results from trials published with significant duplication resulting, for example, across all registered COVID-19 research studies on CT.gov, only 3% had reported results in July 2022 despite 53% being past completion dates.<sup>10</sup>

It is therefore vital to realise that data sharing is fraught with difficulties spanning technical, legal and organisational risks. Even though Open Access is increasingly supported by many, data sharing is less prevalent.<sup>11</sup> During the pandemic, incentives for sharing were high and the dangers of withholding data were equally significant. Yet, there was limited improvement in the wider system that encourages and facilitates data sharing. Instead, the notion of open data is shrouded in complexity and deemed far to

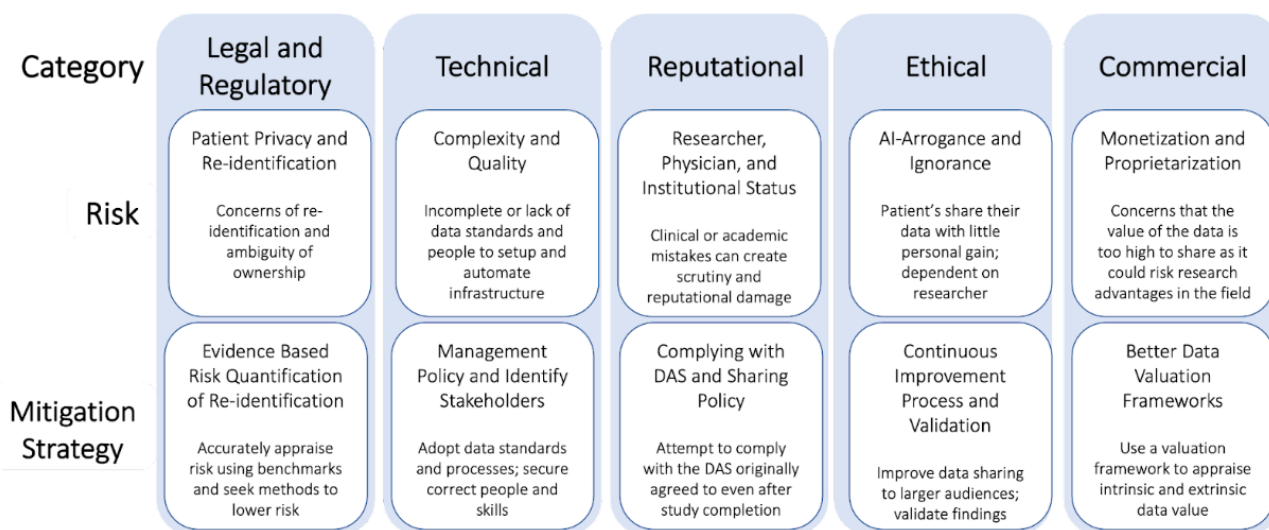


© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

### Correspondence to

Dr Jack Gallifant;  
jack.gallifant@nhs.net



**Figure 1** Data sharing risk and mitigation framework. DAS, data availability statement.

risk. Here, we coin the term Open Data in Appearance Only (ODIAO) defined as the intent of data sharing, but without any actual data sharing occurrence (eg, material or digital data transfers).<sup>12 13</sup>

Data sharing has been debated for many years and across industries, where barriers to distribution have been laid out by The Open Data Institute (ODI), particularly in their 5-year strategy (2023–2028).<sup>14</sup> The ODI notes several vital developments that must be overcome to facilitate data sharing and build stakeholder trust. In addition, this mirrors guidelines from the recent NIH DMS plan, which focuses on improving safe data management and its sharing.<sup>15</sup> Both documents aim to accelerate health research, improve transparency and reduce biases transmitted to downstream tasks. In this review, we explore and summarise key lessons from these two critical reports on data sharing risks and barriers. In order to prevent ODIAO, we sought to harmonise and incorporate these key factors into one overarching framework for data sharing that could be used to deliver the recent NIH initiative (figure 1), addressing each factor in turn.

## LEGAL AND REGULATORY

### Patient privacy and reidentification

The goals of a variety of health and data laws, such as Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health (HITECH) Act, are to protect patient privacy and to create a unified digital infrastructure to improve quality, safety and cost of care.<sup>16</sup> While these initiatives have incentivised practices such as electronic health record (EHR) adoption, there are also penalties and fines for breaches and patient identification<sup>17</sup> that are largely used as reasons not to share data. The risk of reidentification is frequently dependent on knowing pieces of information about a patient outside of bounds of the deidentified data. This includes

other publicly available dataset or personal knowledge. A recent study showed that by using publicly available newspaper data to match names to anonymised patient records in statewide hospital data 28% of names in Maine and 34% of names in Vermont were able to be uniquely matched to one hospitalisation. After redacting the same data to HIPAA Safe Harbor standards the linkage rate was reduced to 3.2% and 10.6% reidentification for Maine and Vermont, respectively.<sup>18</sup> The linkage of hospital data poses privacy risks because it allows previously unknown information within the hospitalisation record including other patient diagnoses to be known such as mental health, addiction or disabilities. Another key example may be an uncommon patient diagnosis code currently onward where a person other than any healthcare practitioner overseeing the patient's care could reference the patient by their diagnosis and then correctly identify the patient by only searching for the patient's International Classification of Diseases (ICD) diagnosis code. The latter example is a violation of protected health information (PHI) practices under the HITECH act and represents the most common cause for a HIPAA breach known as 'data snooping'. Rare disease ICD codes may also be considered quasi-identifiers when combining data with patient forums.<sup>19</sup>

In most cases of deidentified medical resources, a potential data consumer must request access to the database and complete ethical research conduct certifications. Although reidentification cannot be completely mitigated, it is worth considering the possibility of identifying a person's health information without deidentified research data at all. For example, if a person is active on a public patient disease forum and states their disease (ie, myasthenia gravis), general field of work (ie, accounting) and geotagged to their city (ie, Boston). Cross-referencing these data with public records and social media may be enough to reasonably infer information on the person

without deidentified research data at all. In instances such as the above example outlined, privacy may be entirely reliant on the deniable plausibility of being any single individual, known as k-anonymity privacy. This example highlights that with or without 'identifiable' health record data, individual's health data can be vulnerable to wide-scale reidentification using data shared directly by individuals 'consensual', shared via data brokers, or found in the 'public domain'.

### Evidence-based risk quantification of reidentification

The two main components used to quantify risk are the probability and severity of the event. In order to approximate the quantification of the true risk of reidentification factors outside of the data itself need to be considered. The first is the motivation for reidentifying the research dataset. We argue the incentive is lower to breach data for data that is able to simply be requested. In this way, research data are often different from breaching commercial data for usage, such as financial fraud and identity theft. Before being granted access to a research dataset, the user requesting access typically must accept the institution's data use agreement (DUA). The DUA is linked to information about the user including identifiable information and specifies the intended purpose for the data and how it may not be used. DUAs most common term and condition is to make no attempt to learn the identity of any person or establishment within the data, and sanctions for violating the DUA is considered a felony with charges such as imprisonment or fines (the National Center for Health Statistics is imprisonment up to 5 years and US\$250 000 fine).<sup>20</sup>

While reidentification of deidentified does pose a risk to patients, this risk is often systematically overestimated and confused with data exfiltration. In a systematic review of healthcare data reidentification, 14 studies were identified, and 2 studies had been deidentified using standards-based methods.<sup>21</sup> Interestingly, of the 14 reidentification studies, 11 were carried out by researchers, 2 were informed court judgements and 1 by a journalist supporting our claim that reidentification and data exfiltration are commonly conflated and confused. Within one of study standards-based methods commissioned by the US Department of Health and Human Services, it found that only 0.013% of the records could be reidentified, while the other study in the UK used survey data that could only be obtained under very strict confidentiality conditions to reidentify information (that would violate a DUA). Another publication that analysed motor vehicle accidents (MVA) specifically due to newspaper coverage found that even when targeting this specific patient population. The data analysed from the Buffalo, NY area found that by cross-referencing seven indirect identifiers 0.88% of the MVA patients were able to be reidentified compared with the 0.0017% of all database patients.<sup>22</sup> While this difference in patient populations represents a huge increase in relative risk of reidentification, it is worth noting that consideration of both having (1)

stricter deidentification standards in more easily identifiable subpopulations such as MVA and rare disease (and further verified by statistical expertise where possible) and (2) how publicly available information is reported in outlets such as newspapers. One publication found that by knowing 15 demographic attributes, 99.98% of the population could be reidentified.<sup>23</sup> However, not all attributes were found to have the same level of uniqueness where attributes such as race, gender and citizenship did not give a considerable lift to the reidentification accuracy, additionally, highly unique, and therefore, identifying pieces of information such as the full date of birth and zip code were included in the analysis would not satisfy HIPAA Safe Harbor standards of deidentification. Finally, we acknowledge that reidentification efforts outside of research activity would be less likely to be published in the first place, particularly if the goal is for information gain to be used in an advantageous or illicit way.

Within an organisation, each person who works with data has responsibility to understand the data risks and mitigation through proper HIPAA training and data transfer processes. While the HHS has outlined the methods for deidentification standards as Safe Harbor or expert deidentification,<sup>24</sup> what constitutes 'very small risk' rightfully remains subjective. Such questions an organisation may need to ask are: Who is responsible for this risk assessment and mitigation? How is this risk evaluated? Do these individuals correctly estimate the risk associated with data breach? Do they consider the use of this data to increase the likelihood of data breach? Organisations will view these risks differently; however, by standardising the approach to each of these questions, a systematic approach can be repeatedly performed. Thus, allowing more accurate depiction of risk that can be more readily quantified with the intention of more frequent data sharing in the future. The NIH DMS mandates data sharing being conducted under their funding; the development of an organisational approach to risk monitoring is a necessary accompaniment that would build trust and prevent ODIAO.

### TECHNICAL Complexity and quality

Due to the rapid adoption and large-scale deployment of digital technology in our society, Big Data and related analytics have become ubiquitous for supporting decisions and operations. However, the volume, variety, velocity and veracity of new data bring new complexities and concerns on information quality. A typical example is the ongoing challenge of data sharing in cities, which are frequently used and combined within healthcare research. Thanks to the invention of low-cost sensors, cloud computing and personal digital devices, cities nowadays enjoy rich information resources to assist data-driven decision-making and automated operation. However, the digitisation of urban systems and internet society bring new social and technical complexities. New

information types and data formats create technical and social complexities for sharing data. One practical challenge is a lack of computing expertise for properly creating, managing, processing and exchanging data. A previous study investigating data landscape in US cities reveals a significant variety and disparity of data formats in city open data, particularly the structured, tabular data (eg, less than 40% of city data in Boston are in tabular format).<sup>25</sup>

Successful data diplomacy practice only starts from data sharing but completes with effective information integration and implementation. Even though multiple parties are willing to share data, a lack of standard in definition and classification may still cause data integration failures, preventing greater value creation. One example is the digital building permitting system in US cities. One recent study investigated building permit data in eight major US cities and found various terminology and classifications, although the data are publicly available and report similar information.<sup>26</sup> Such a lack of data standards brings difficulties in quality evaluation and integrated analytics across multiple cities. Beyond the technical barriers, additional non-technical concerns involve unexpected social impact. For example, several cities previously have published aggregated academic performance data by school districts without personal-identifiable information. However, such information-sharing caused concerns on creating discriminations and biases towards specific neighbourhoods, particularly on the local housing value estimation and property market appreciation. Such unwanted consequences to certain communities and population groups create additional complexities in data sharing and information publication.

### Data management policy

To cut through the data complexity and quality issues in data sharing development of a data management policy and identifying the correct stakeholders is crucial. The goal of a data management policy is to deliver the right data to the right user at the right time with the lowest possible cost and friction. The data management policy outlines considerations such as what data standards are followed, where the data will be stored, what requirements there are to access the data, how data will be accessed, what the time frame of the data is from, how to join the data and schema information, and include data descriptions and data dictionaries. For example, the Medical Information Mart for Intensive Care (MIMIC-III) database is accessible through both Google Cloud Platform and Amazon Web Services, is accessed through PhysioNet, Collaborative Institutional Training Initiative (CITI) training is required, and date fields and filters are stated within the data itself. An important component of the data management plan to highlight is adoption of data standards. Data standards are documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use and management of data.<sup>27</sup> By implementing data standards prior to

collecting data where possible, the amount of data governance and structure can be reduced by avoiding remapping and standardising data at a later time. Currently, the data cleaning stage of a project takes the most amount of time, by implementing data standards code, presentations, publications and information quality can be reproduced and validated in less time than without data standards. A notable data standardisation is the Fast Healthcare Interoperability Resources, which is a standard for healthcare data exchange that addresses many aspects of health from diagnostics and medications to claims and genomics.<sup>28</sup>

### Creating a village mindset

There are several key stakeholders that must work together with a 'village mindset' in order to make data sharing possible.<sup>29</sup> Here, we outline specific roles, but a single person may represent multiple skill sets and contribute to the data diplomacy ecosystem. Generally for an institution to make a data transfer, there will be approval and strategy, legal, technical and considerations. Our aim is to help organisations accurately identify gaps in people and skills that, if bridged, will facilitate more standardised and swifter data sharing. A data engineer is able to extract the data from the source system, create data quality metrics, filter and aggregate the data, and set up the means in which it will be transferred. For small data sizes, the transfers could be set up through simple cloud storage sharing (ie, box). Larger datasets may require cloud computing (S3, Redshift, Blob, etc) and a Secure File Transfer Protocol or managed roles to control access and port over data. The chief data officer (CDO) is a senior executive responsible for the stewardship, utilisation and governance of an organisation's data. Typically, the CDO's approval is required to sign off on entering into a data sharing agreement. An organisation may also have a chief information officer (CIO) instead as the reviewer. The chief privacy officer (CPO) is responsible for developing, implementing and maintaining policies designed to protect employee and patient data from unauthorised access. Such policies could involve technical access controls to only certain internal personnel or could be non-technical such as running HIPAA and PHI training at regular cadences. General and legal counsel will be involved in the approval and may be responsible for running training to explain the nuances in HIPAA policies such as explaining risk differences between covered entities and business associates. The CDO or CIO work alongside the CPO to create a data sharing agreement outlining the possible risks and synergies from the agreement. Once the format and transfer means are agreed on, the data engineer is able to execute on creating the correct dataset and setting up transfer ports. If the data sharing is maintained through a data sharing platform, there will be additional technical personnel involved, such as cybersecurity and site reliability engineers that are not elaborated on in the scope of this work.

While technical advances continue to be made, the complexity of the data being used and the types of agreements being made continue to grow. Clear standards for data sharing must be provided by governing bodies but must also be set locally as well for internal processes. Organisations must involve a wide range of disciplines and backgrounds in this process to maximise the chance of data usage and prevent data siloing that can lead to ODIAO.

## REPUTATIONAL

### Researcher, physician and institutional status

A personal barrier to withholding data can be found in the lack of willingness for errors to be found. What is a completely natural response, however, merely delays the time at which the mistake is uncovered. As failure to replicate results sparks investigation. This is both a waste of time and resources as well as potentially putting patient lives at stake. So, although it may appear that refusal to share data avoids the risk of academic or commercial scrutiny. Refusal to share data does not ultimately protect reputation; it masks issues and impedes discovery, innovation and discourse over time. Retraction Watch, part of the Center for Scientific Integrity, has reported a significant rise in the number of retractions each month,<sup>30</sup> particularly since the COVID-19 pandemic.<sup>31</sup> Echoing the number of high-profile cases of fraudulent research.<sup>32</sup> The distribution of data and code would normalise corrections, improve patient safety and reduce duplication of work that attempts to replicate results.

Duplication of research also carries another risk, data breaches. Data breaches are infrequent but can be significant, affecting a large number of patients. Although, in an open data environment, more data will be public, similar volumes of research will still be conducted. By increasing access to standardised and secure data environments, a higher proportion of research would be hypothetically performed in a regulated and secure setting. This relies on data sharing being appropriately regulated to shift the burden of risk from the researchers to the governing organisation.

### Complying with data availability statement and regulation

The purpose of data availability statements (DAS) is to provide information regarding where the data supporting the findings in a published article can be found, and if and how they can be accessed. These policies are part of a broader movement to encourage open science and data sharing. Depending on the types of data involved, however, there can be tension between the data sharing promoted through DAS and privacy regulations. Qualitative and mixed-methods research, for example, may contain data that is difficult to sufficiently anonymise in order to prevent deductive disclosure.<sup>33</sup> Recent studies have found, though, that many researchers do not comply with what they set out in their DAS, and even that there was not a difference in compliance rates for articles that

have a DAS compared with those that do not.<sup>34 35</sup> Notably, the study found that 80% of corresponding authors did not reply to the contacts for a data request, and of the 20% that did respond, only 50% shared the data. Overall, this yielded a 93% non-response rate or decline to share data.

The General Data Privacy Rule (GDPR) enacted by the European Union (EU) gave stronger privacy protections to individuals by requiring stronger consent and providing new rights to be forgotten and for data portability. While there were initial concerns that the GDPR would impede scientific data sharing, the final version included exemptions that supported data sharing for scientific research.<sup>36</sup> With more complex collaborative arrangements for scientific data sharing, though, there can be a need to establish clearer roles in the data sharing networks under the GDPR.<sup>37</sup> A 2021 report found that the GDPR was having a negative impact on oncology and other types of health research, in part because it hampers the sharing of data outside of the EU, thus making it more difficult to share data as part of international collaborative health research.<sup>38 39</sup> Therefore, both correct interpretation of GDPR and identification of stakeholder responsibilities is necessary.

The new NIH DMS Policy requirement will combine the expectations of proper data management and sharing by formalising the plan as part of its application process. This includes considerations for: describing the data types; related tools, software and/or code; data standards; data preservation, access and associated timelines; and access, distribution or reuse considerations.<sup>15</sup>

## COMMERCIAL

### Monetisation and proprietarisation

In the last decade, data based startups, academic spin-outs turned companies, and patents on data processing have become more common<sup>40</sup> and data have been referred to as the new oil, by Clive Humby as early as 2006, in the digital and information age. Data sharing can be seen as a risk to both monetisation and proprietarisation if the data asset is core to the research or product. We argue that although data in itself may have some inherent value, it is a building block to higher value insights requiring context to become information, meaning to become knowledge and insight to become wisdom. Each of these stages to transform data into solutions to real-world problems and helping patients requires personnel with specialised technical and subject matter expertise.

For others, the prospect of making institutional data accessible to those outside of the organisation will allow others to benefit from the data, and this may be viewed as the loss of an asset without compensation. This is compared other groups that charge researchers, institutions and industry licensing fees for data access. In the same study analysing DAS, some corresponding authors proposed or expected coauthorship for use of the data, representing an expectation of proprietarisation on

secondary analysis.<sup>35</sup> These types of expectations in the research system make it difficult for analysis to refute original claims or propose divergent hypotheses.

### Better data valuation frameworks

Current data valuation approaches for institutions and organisations are ambiguous and vague at best and non-existent at its worst. The idea that the value of data solely resides in what another party would be willing to pay is reductionistic and typically represents only a small fraction of the data's value. Data value would be better valued by its ability for the data to optimise an operation or act in support of a larger product or process.<sup>41</sup> For example, a hospital may want to optimise hospital bed capacity and use parameters such as transfers, unscheduled admissions and unoccupied beds to derive an optimisation model.<sup>42</sup> Making these data used to create the optimisation model available on request through a DAS does not automatically mean that the data will be insightful, generalisable, or actionable to other hospitals for their gain. Finally, by making data available through a DAS, it does not lower significant barriers such as highly specialised personnel, team size, legal assistance and cloud compute costs that usually make data monetisation and proprietarisation possible.

When valuing a data asset, instead of assigning an absolute nebulous worth to the data, it is best to contextualise the data asset in terms of its utility for the problem trying to be solved.<sup>43</sup> Factors to include in data valuation may consist of the data's: strategy, features, size, granularity, quality, standards and processes to create a more meaningful understanding of data utility.

Organisations and researchers must find a middle ground where they are rewarded for efforts in dataset collection, curation and storage yet still maximise access to data that has the potential to improve patient outcomes. The maturation of DAS' and guidelines such as the NIH DMS will help to safeguard the inevitable competition of monetisation through scarcity and beneficial impacts of data sharing.

## PSYCHOLOGICAL

### AI arrogance and ignorance

The current system means that the risk for sharing one's data is high, with little personal gain. Despite the fact these risks are real to the institution, the failure to disclose data does not eliminate the risk; it merely transfers the risk from the institution to the patients being treated based on the research. Thus, those who we claim to be helping must carry the risk for our own arrogance and ignorance, which may be worse than fatal, where one's data may worsen the outcomes of another human being who 'does not look like you'. This problem can be further exacerbated by reasoning that Artificial Intelligence (AI) methods such as synthetic minority oversampling technique (SMOTE) will simply 'fix' issues such as sex and race data imbalances. AI has introduced new effective

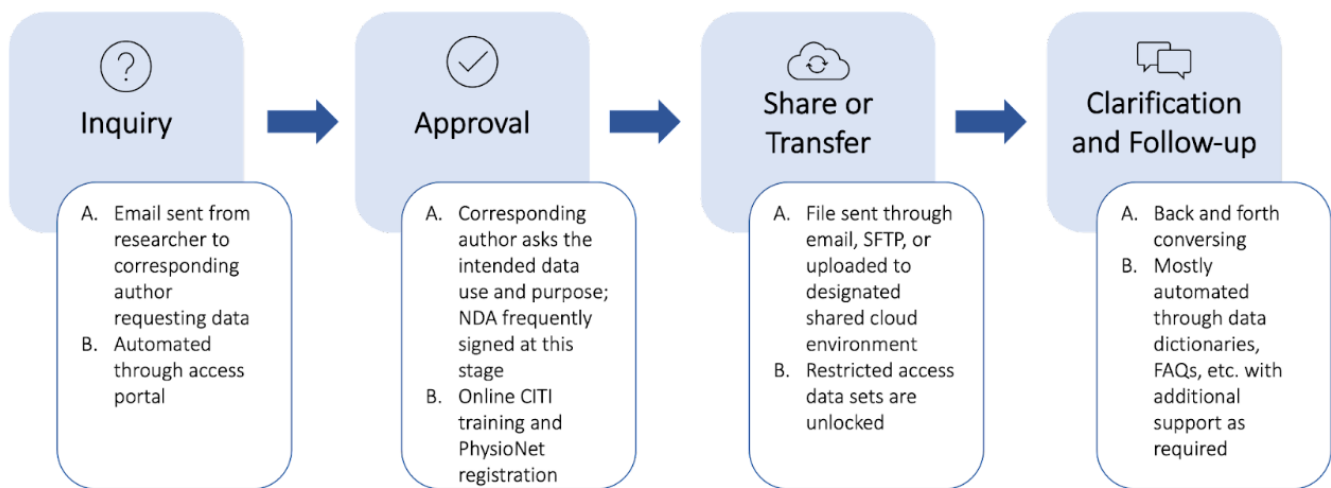
methods, such as SMOTE that can forward medical and social issues, but is not a 'cure all' and is instead a specific methodological tool.<sup>44</sup> Current popular interpretation methods such as local interpretable model-agnostic explanations and SHapley Additive exPlanations have respective limitations such as model reduction to an alternative localised linear or probability values for covariates that are in reality collinear.<sup>45</sup> These limitations are not a sole reason to discard them, but be thoughtfully instead of blindly executed. Methods that intersect AI and causal frameworks that perform counterfactual scenarios about outcomes based on attributes should not be implemented indiscriminately on features conditional on each other.<sup>46</sup> For example, if you wanted to understand a survivor expectancy of a male patient if instead they were female, other attributes such as occupation, income level, age and race would need to be considered holistically.

Historically, tools and software used for research are specified in publications, but code sharing is newer and less frequently incorporated as part of the publication or supplement. As AI and coding are linked, so is AI arrogance and lack of code sharing and transparency. Much like the DAS, code is available on request. While the true availability of the data outlined in DAS statements has begun to be researched, code sharing is not specifically well researched and is likely more researched in specifically computational journals.<sup>47</sup> While tools and software may by nature use Graphical User Interfaces (GUIs) that cannot be automatically reproduced by being run, coding scripts are. While code sharing is possible through Git and providers such as GitHub and GitLab there are legal, technical and reputational risks associated with sharing source code. These can span from how deidentification is conducted to critiques ways the code is more methodically robust, scalable or elegant (few lines of code). By turning the research focus back to patient centricity, the risks posed by code sharing are smaller compared with the issues of non-reproducibility and model improvements.

### Continuous improvement process and validation

A discontinuous and stochastic approach dominates current quality improvement, however, a mindset shift towards a data-centric and systems-based methodology should be leveraged in the future. In order to make data sharing a more frequent reality that acts in service of the patient, incremental change at the organisation, researcher and data set level are required. A continuous improvement process for data sharing means iterating on the parts of the process that cause failure. It is distinct from the data management plan; while a data management plan is created before or during data sharing and primarily completed once the data is shared, a continuous improvement process is cyclical. While a continuous improvement process has technical aspects, it is driven by considerations of an organisation to serve both the patient and research community.<sup>48</sup>

Typically, the data sharing process begins with how a data sharing inquiry is received and to whom, the



**Figure 2** Data sharing process with manual and automated scenarios, A and B, respectively. Non-Disclosure Agreement, NDA; CITI, Collaborative Institutional Training Initiative; SFTP, Secure File Transfer Protocol; FAQs, Frequently Asked Questions.

approval process, the data transfer and/or sharing, and clarification and follow-up support. The goal of a continuous improvement process for data sharing means first designing with data in mind and iterating on the pain points for greater data dissemination.<sup>49</sup> Figure 2 illustrates two possible process flows for a four-step data sharing process, A and B. Scenario A represents what the data sharing process looks like without any online data repository or portal and scenario B represents where a repository or portal solution has been implemented. From the inquiry to clarification and follow-up scenario A has many more communication and back and forth touchpoints between the corresponding author and the researcher making the request. Scenario B outlines the type of data sharing process that is possible when a continuous improvement process is implemented with the patient and research community in mind.

From these two scenarios, we can glean that through a continuous improvement process there are opportunities to potentially automate and reduce the time and effort exerted to share data (figure 3). A continuous improvement process laid out by an institution may consist of multiple aims such as to use a trusted research database portal, begin adopting data standards of the field prior to data collection of an experiment, and incorporate a deidentification requirement for project completion with the intent of data sharing. By placing data sharing as a goal to be met in service of the patient and research community, it is less likely to be considered an after thought or extra work with low incentivisation for the researcher. A continuous improvement process is not seen as complete, as new needs arise whether making the data sizes more accessible or creating documentation for frequently asked questions about the data set, the process is aimed to give the best possible experience in sharing and understanding the data. An exemplar system that lifts the onus of data sharing from the researcher entirely is MIMIC. The MIMIC data are accessible via PhysioNet, where data

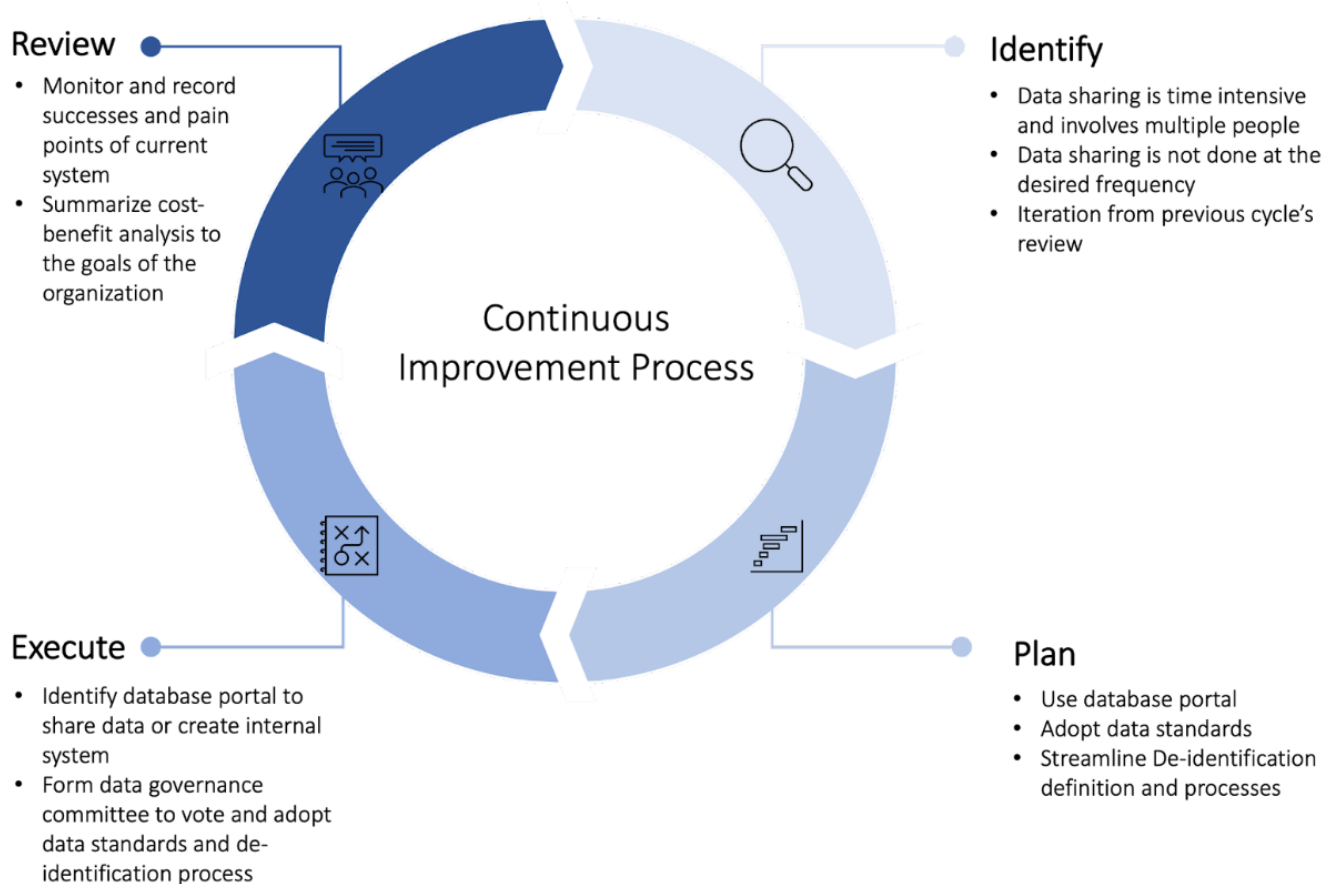
sets are categorised as open, restricted or credentialed. For restricted data sets, including the latest version of MIMIC, CITI training must be completed, user information and completing the DUA are required. Additionally, data dictionaries, release notes specifying incorrect data and subsequent corrections, and directions for how to join commonly created data views are documented for MIMIC.<sup>50</sup>

Data sharing currently emphasises the ability to garner better scientific reproducibility, but validation is equally if not more important. From a treatment perspective, it is imperative to prove clinical efficacy such as AI enabled treatment recommendations created from longitudinal analysis of demographic, symptom and vital sign data. By putting the patient first, AI then refocuses itself as a tool, where clinical safety and efficacy supersedes importance of AI interpretability and explainability.<sup>51</sup> Where AI is used to lead to treatment enhancement indirectly in medical imaging analysis or organisation of unstructured EHR data, validation of the accuracy of the method, the degree of utility and ability to generalise is where patients can benefit. To mitigate AI research, arrogance and ignorance, goals need to be oriented so there is a direct relationship from the patient providing their data to improvements in health outcomes.

The recent NIH initiative forces the sharing of such data and, thus, we hope, a change in mindset that promotes humility and transparency. The development of continuous and systematic approaches to quality improvement are a beneficiary of such a mindset. Further, it shares the same psychological sentiment that drives data sharing and would discourage ODIAO.

## CONCLUSION

There is a growing acknowledgement that data sharing is likely in patients' best interest; however, we identified five key barriers that can oppose data sharing and



**Figure 3** Continuous improvement process for more ubiquitous data sharing.

lead to ODIAO. A mindset shift is required to prioritise patient-centred research in a system where data are a valuable asset and mitigate real patient privacy risks that need to be quantified. In order to realise the benefits of data sharing while navigating such risks, the NIH 2023<sup>3</sup> mandate must be actively supported by a village mindset that cultivates the talents of all stakeholders. The postpandemic world needs data sharing to become a cornerstone of health research, to safeguard against the implementation of harmful treatments and algorithms. Moreover, to encourage public data sharing, there must be incentives driven from the bottom up starting with the patients themselves. The NIH DMS is a valuable start to this and strongly opposes ODIAO.

#### Author affiliations

<sup>1</sup>DBT Labs, Boston, Massachusetts, USA

<sup>2</sup>Department of Critical Care, Guy's and St Thomas' Hospitals NHS Trust, London, UK

<sup>3</sup>Department of Urban Planning and Design, Tsinghua University, Beijing, China

<sup>4</sup>New York University Marron Institute of Urban Management, New York, New York, USA

<sup>5</sup>University of Texas Southwestern Medical Center, Dallas, Texas, USA

<sup>6</sup>Southern Methodist University, Dallas, Texas, USA

<sup>7</sup>Heidelberg Institute of Global Health, Heidelberg University, Heidelberg, Germany

<sup>8</sup>Department of Medicine, Instituto Politécnico Nacional, Ciudad de Mexico, Mexico

<sup>9</sup>Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, California, USA

<sup>10</sup>Department of Radiology, Emory University, Atlanta, Georgia, USA

<sup>11</sup>UNICEF, New York, New York, USA

<sup>12</sup>Laboratory for Computational Physiology, Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA

<sup>13</sup>Division of Pulmonary Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

**Twitter** Judy Gichoya @judywawira, Uyen Kim Huynh @ukhuynh and Leo Anthony Celi @MITCriticalData

**Contributors** HW: conceptualisation, methodology, writing—original draft preparation, writing—reviewing and editing. JG: conceptualisation, methodology, writing—original draft preparation, writing—reviewing and editing. YL: conceptualisation, methodology, writing—original draft preparation, writing—reviewing and editing. APR: conceptualisation, methodology, writing—original draft preparation, writing—reviewing and editing. CV: methodology, writing—reviewing and editing. NM: methodology, writing—reviewing and editing. JG: methodology, writing—reviewing and editing. UKH: methodology, writing—reviewing and editing. LAC: conceptualisation, methodology, writing—reviewing and editing, supervision.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data sharing not applicable as no datasets generated and/or analysed for this study.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.



## ORCID iDs

Jack Gallifant <http://orcid.org/0000-0003-1306-2334>

Judy Gichoya <http://orcid.org/0000-0002-1097-316X>

Leo Anthony Celi <http://orcid.org/0000-0001-6712-6626>

## REFERENCES

- Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018:160018..
- Liu J, Carlson J, Pasek J, *et al.* Promoting and enabling reproducible data science through a reproducibility challenge. *Harvard Data Science Review* 2022.
- NOT-OD-21-013: final NIH policy for data management and sharing. Available: <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html> [Accessed 13 Feb 2023].
- Parkes DC. Building a more robust data science, toward a more robust future. *Harvard Data Science Review* 2022.
- Martone M, Nakamura R. Changing the culture on data management and sharing: getting ready for the new NIH data sharing policy. *Harvard Data Science Review* 2022.
- Tabak L, Jorgenson L, Martone M, *et al.* Conversation with Dr. Lawrence Tabak and Dr. Lyric Jorgenson on the NIH perspective on data sharing and management. *Harvard Data Science Review* 2022.
- Lowenberg D. Recognizing our collective responsibility in the Prioritization of open data Metrics. *Harvard Data Science Review* 2022.
- Smith PC, Mossialos E, Papanicolas I, *et al.* *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge University Press, 2010.
- Watson C. Rise of the preprint: how rapid data sharing during COVID-19 has changed science forever. *Nat Med* 2022;28:2–5.
- Dron L, Kalatharan V, Gupta A, *et al.* Data capture and sharing in the COVID-19 pandemic: a cause for concern. *Lancet Digit Health* 2022;4:e748–56.
- Bjaalie JG, Goble C, Sansone S-A, *et al.* Perspectives on data sharing and the new NIH policy from the European Union. *Harvard Data Science Review* 2022.
- Gallifant J, Zhang J, Del Pilar Arias Lopez M, *et al.* Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias. *Br J Anaesth* 2022;128:343–51.
- Watson C. Many researchers say they'll share data - but don't. *Nature* 2022;606:853.
- ODI strategy 2023–2028. Available: <https://www.theodi.org/about-the-odi/odi-strategy-2023-2028/> [Accessed 03 May 2023].
- Writing a data management & sharing plan [Data Sharing]. 2022. Available: <https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-DMS/writing-a-data-management-and-sharing-plan> [Accessed 08 Aug 2022].
- Rights (OCR) O for C. HITECH act enforcement interim final rule. HHS.gov; 2009. Available: <https://www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule/index.html> [Accessed 09 Aug 2022].
- Office for Civil Rights. HIPAA violation cases - updated [HIPAA J]. 2022. Available: <https://www.hipaajournal.com/hipaa-violation-cases/> [Accessed 09 Aug 2022].
- Yoo JS, Ra Thaler A, Thaler L, *et al.* Risks to patient privacy: a re-identification of patients in Maine and Vermont statewide hospital data [Technol Sci]. 2018. Available: <https://techscience.org/a/2018100901/> [Accessed 08 Aug 2022].
- Gow J, Moffatt C, Blackport J. Participation in patient support forums may put rare disease patient data at risk of re-identification. *Orphanet J Rare Dis* 2020;15:226.
- Data access - data user agreement. 2020. Available: [https://www.cdc.gov/nchs/data\\_access/restrictions.htm](https://www.cdc.gov/nchs/data_access/restrictions.htm) [Accessed 10 Aug 2022].
- El Emam K, Jonker E, Arbuckle L, *et al.* A systematic review of re-identification attacks on health data. *PLOS ONE* 2011;6:e28071.
- Janmey V, Elkin PL. Re-identification risk in HIPAA de-identified datasets: the MVA attack. *AMIA Annu Symp Proc* 2018;2018:1329–37.
- Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019;10:3069.
- Rights (OCR) O for C. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule [HHS.gov]. 2012. Available: <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html> [Accessed 11 Aug 2022].
- Barbosa L, Pham K, Silva C, *et al.* Structured open urban data: understanding the landscape. *Big Data* 2014;2:144–54.
- Lai Y, Kontokosta CE. Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. *Computers, Environment and Urban Systems* 2019;78:101383.
- US EPA O. Learn about data standards. 2015. Available: <https://www.epa.gov/data-standards/learn-about-data-standards> [Accessed 12 Aug 2022].
- Index - FHIR V4.3.0. Available: <https://www.hl7.org/fhir/index.html> [Accessed 11 Aug 2022].
- Borgman CL, Bourne PE. Why it takes a village to manage and share data. *Harvard Data Science Review* 2022;4.
- Oransky I. Retractions are increasing, but not enough. *Nature* 2022;608:9.
- Moradi S, Abdi S. Pandemic publication: correction and Erratum in COVID-19 publications. *Scientometrics* 2021;126:1849–57.
- Piller C. Blots on a field? [Science]. 2022. Available: <https://www.science.org/content/article/potential-fabrication-research-images-threatens-key-theory-alzheimers-disease>
- Tsai AC, Kohrt BA, Matthews LT, *et al.* Promises and pitfalls of data sharing in qualitative research. *Soc Sci Med* 2016;169:191–8.
- Federer LM, Belter CW, Joubert DJ, *et al.* Data sharing in PLOS ONE: an analysis of data availability statements. *PLOS ONE* 2018;13:e0194768.
- Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *J Clin Epidemiol* 2022;150:33–41.
- Meszaros J. The conflict between privacy and scientific research in the GDPR. 2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC); 2018:1–6
- Becker R, Thorogood A, Bovenberg J, *et al.* Applying GDPR roles and responsibilities to scientific data sharing. *International Data Privacy Law* 2022;12:207–19.
- ALLEA (European Federation of Academies of Sciences and Humanities), FEAM (Federation of European Academies of Medicine), EASAC (European Academies' Science Advisory Council). International sharing of personal health data for research. DE: ALLEA; 2021.
- Gourd E. GDPR obstructs cancer research data sharing. *Lancet Oncol* 2021;22:592.
- Teare G. How seed funding has exploded in the past 10 years [Crunchbase News]. 2021. Available: <https://news.crunchbase.com/venture/seed-funding-startups-top-vc-firms-a16z-nea-khosla/> [Accessed 12 Aug 2022].
- Tufiş M, Boratto L. Toward a complete data valuation process. challenges of personal data. *J Data and Information Quality* 2021;13:1–7.
- Ravaghi H, Alidoost S, Mannion R, *et al.* Models and methods for determining the optimal number of beds in hospitals and regions: a systematic scoping review. *BMC Health Serv Res* 2020;20:186.
- Infonomics - Monetize, manage & measure information [Gartner]. Available: <https://www.gartner.com/en/publications/infonomics> [Accessed 12 Aug 2022].
- Takaoka K. AI implementation science for social issues: pitfalls and tips. *J Epidemiol* 2022;32:155–62.
- Molnar C. Chapter 9 local model-agnostic methods [Interpretable Machine Learning]. Available: <https://christophm.github.io/interpretable-ml-book/local-methods.html> [Accessed 12 Aug 2022].
- Lundberg I, Johnson R, Stewart BM. What is your Estimand? Defining the target quantity connects statistical evidence to theory. *Am Sociol Rev* 2021;86:532–65.
- Cadwallader L, Papin JA, Mac Gabhann F, *et al.* Collaborating with our community to increase code sharing. *PLOS Comput Biol* 2021;17:e1008867.
- Fecher B, Friesike S, Hebing M. What drives academic data sharing. *PLOS ONE* 2015;10:e0118053.
- First, design for data sharing [Nature Biotechnology]. Available: <https://www.nature.com/articles/nbt.3516> [Accessed 12 Aug 2022].
- Johnson A, Pollard T, Mark R. MIMIC-III clinical database. *PhysioNet* 2015.
- Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021;3:e745–50.