**Nature Digital Medicine**

**Collaborating without Sharing—Decentralizing Data Availability**

Mary-Anne Hartley M.D., Ph.D; Martin Jaggi Ph.D

**ABSTRACT**

**As medical data becomes ever bigger and more granular, so too does its potential power, value, risk, and the practical constraints of sharing it. Increasingly, our copy-paste mentality of centralizing static data duplicates is proving unsustainable and creates ethical concerns on ownership, privacy, and consent. Several validated methods exist to collaboratively learn from data without sharing it at all: by distributing algorithms to the source of the data (rather than centralizing the data to the source of the algorithm). We summarize the rationale for integrating this technology into collaborative clinical research and invite contributions to an open-source web-based platform to guide a needs-based, community-driven design.**

By 2025, global data is estimated to reach 168 zettabytes, with a projected 25 petabytes of genomic data added annually (approximately equivalent to 10 trillion pages of text).

As the bigness of big data becomes ever bigger and more granular, so too does its potential power, value, risk, and the practical constraints of sharing it. Increasingly, our copy-paste mentality of centralizing static data duplicates is proving unsustainable and unrealistic.

Clinical data is often not shared for a range of well-considered reasons, such as concerns over patient privacy, intellectual property, and laboriously necessary ethicolegal constraints. The vast majority of medical data is thus siloed by institution or in the ownership of single researchers. This not only fragments the statistical power of predictive models but creates an accessibility bias, where accuracy becomes inequitably distributed to those who have the resources to overcome these constraints or collect big enough datasets of their own **[1]**.

Even when data is shared, duplicating it into a read-only repository uncouples it from its source, failing to benefit from subsequent data cleaning, harmonization efforts, or real-time updates from ongoing collection, creating significant delays in sharing and utility. Then, there are the bigger issues of how to guarantee privacy and control over patient-level consent. The act of removing names and dates of birth are obfuscation techniques of yesteryear. New machine learning approaches can more easily identify unique individuals, which is the first step of re-identification and puts many previously published datasets at retrospective privacy risk. Indeed, data privacy is an evolving concept that erodes with computational advances (remember this **[2]**?), and so, many feel that the current standard of making static patient-level data copies available in perpetuity is no longer responsible.

Beyond privacy and sustainability, there are further important concerns about the sovereignty and ownership of a central repository. Who should own it? Where should it be situated? Who pays for the storage? How does one update it? Is it gate-kept for content to manage costs and avoid the creation of ‘data dumpsters’? If so, what is considered ‘worth’ curating? Who would do that (and how)? Who can download from this repository (i.e., *again*, copy-pasting the data…)? Everyone? Or just ‘researchers’? Who is a ‘researcher’? And perhaps most importantly, how can we control that the data will be used ethically and securely once downloaded?

These newly emerging risks, costs, and ethical concerns further disincentivizes collaboration despite the increasingly obvious responsibility to do so (c.f. Ebola, Zika, COVID, Monkeypox etc.). The issue is particularly sensitive for clinical trials, where participants put their health (and lives) at risk to advance science. It is then expected that researchers maximize this benefit of advancing science without putting participant privacy at risk. These responsibilities, however, are often still trumped by the *Fear of Getting Scooped*, where researchers have strong concerns about sacrificing their intellectual property (IP) before publishing **[3]**.

While journals have made an effort to clarify data accessibility by incorporating availability statements into publications, a recent study on the authenticity of the claims made in these statements revealed that over 90% of promises of “sharing upon reasonable request” were likely empty **[4]**.

Collaborative learning is an emerging approach offering a solution to many of these concerns by allowing copy-paste-free data analysis and model building from distributed silos without sharing any original data **[5]**.

Here, instead of sharing the data centrally to train a model, the model is shared directly with the distributed data sets. The model updates are then securely communicated and aggregated into a common shared model. Thus, flipping the archetype of *public data—private model* to *public model—private data*.

**Figure 1** depicts the two forms of collaborative learning, defined by how the model updates are communicated and aggregated: either coordinated by a central server (federated learning, **Fig 1a**) or delegated to individual data owners (decentralized learning **Fig 1b**).

**Note to editor:** These are animated gifs (which may not work in a PDF generated by the submission process)

If they are not animated, please see the figures here: <https://epfml.github.io/disco/#/information>

|  |  |
| --- | --- |
| **a.** Federated Learning | **b.** Decentralized Learning |
|  | **Diagram  Description automatically generated** |
| **Figure 1.** The different communication strategies of collaborative learning where models are shared to each user and then updates are either communicated to **(a)** a central server for aggregation or **(b)** directly between users. From [https://epfml.github.io/disco/#](https://epfml.github.io/disco/) | |

Both federated and decentralized learning have become well-established in over 20’000 recent scientific publications and are already deployed in industry at scale. These techniques reliably replicate results learned from centralized, shared data. The models learned can range in complexity from logistic regression on tabular data to deep neural networks on any data modality, such as images, text, or sound.

This is particularly practice-changing, as it provides an unprecedented level of control over the ownership, use, and sovereignty of the data which can be updated in real-time. It can also guarantee the ethical intended use of secondary analyses and protect IP, where users can restrict the type of question asked of the data and even the type of model used. This could allow data availability statements in research articles to directly include and enforce the ethical limits of secondary data use (rather than those limitations being a barrier to sharing).

An example of such a statement could be:

|  |
| --- |
| *‘****All******data*** *are available for* ***collaborative learning*** *by* ***any models*** *which seek to predict the outcome of ‘Ebola diagnosis’, as stipulated in the ethical use of this data’.* |

In this example, the owner allows others to use their data to train models predicting *Ebola diagnosis* from all variables, like haemoglobin. Previously, they might have been hesitant to release a copy of the data in a central repository for reasons of IP (e.g. they had ongoing work predicting haemoglobin), or ethical constraints (fearing others may attempt to predict the identity of patients).

In the common scenario that the data is proprietary and training a model on any of the data would infringe IP, the statement could be restricted to only allow validation (testing) to ensure reproducibility.

|  |
| --- |
| *‘The* ***test data*** *is available for* ***collaborative validation*** *by* ***the specific model provided*** *in this study’.* |

However, in these cases, benchmarking should be encouraged, where the owner allows *other* models to be validated on their test set to promote comparative model auditing.

|  |
| --- |
| *‘The* ***test******data*** *is available for* ***collaborative validation*** *on* ***any models*** *which seek to predict the outcome of ‘Ebola diagnosis’, as stipulated in the ethical use of this data’.* |

While this approach may seem to inherently answer all the challenges of data-based collaboration, some endure. For instance, interoperability remains just as important, but in collaborative learning, the process of harmonisation is even more delegated to data owners. This could, arguably, encourage the adoption of global standards, especially if interoperability becomes a condition of publication. A natural question that arises, is *how does one even know if there is an interoperability issue, if the data can’t be inspected?* There are ways to address this issue without compromising privacy such as by simply sharing aggregate statistics, or by deploying advanced techniques that identify data shifts for each participant relative to others [6]. This could inform collaborative troubleshooting between data owners and provide feedback on harmonisation for continual improvement on the interoperability of data.

Another issue that remains in collaboratively built models is the potential of data poisoning, where a real or maliciously manufactured data shift can bias insights. Although, there are several well-described approaches to detect and ignore malicious actors with unsuitable data, achieving what is called ‘Byzantine robustness’. This approach bears analogy to model personalization, which is routinely practiced. For instance, if you are learning an aetiology-of-fever model for a population in Burkina Faso, would you want insights from Sweden? In the extreme, purposely ignoring ‘different’ data risks creating an echo chamber not dissimilar to the currently siloed status quo of fragmented locally learned models, and so, personalisation should be adapted to the intended use-case. Several well-described approaches exist in distributed settings and mirror the quality of models achievable with centralized data.

Collaborative learning also brings with it several novel computational issues and vulnerabilities that need to be stress-tested in real-world use. For instance, the reliance on a central server in federated learning places the intellectual property of the model and updates at the discretion of a central entity, representing a single point of failure. This makes decentralized learning more attractive, as ownership is more fairly distributed, and users get to take more fine-grained decisions on which type of analysis they will allow on their data.

Malicious actors may also use model updates to perform attacks that obtain information about the unshared data. Thus, secure aggregation techniques are required, such as encryption, multi-party computation (where no single user has all the information necessary to reconstruct the data), differential privacy (where carefully calibrated noise is injected into the updates to ensure privacy whilst maintaining model performance), or simply restricting the number and type of questions asked.

We invite readers inspired to contribute to and validate this promising technology to join us at **DISCO** ([https://epfml.github.io/disco/#](https://epfml.github.io/disco/)): an open-access, open-source **DIS**tributed **CO**llaborative web-based learning platform, offering users a means to collaboratively build machine learning models without sharing any original data. The modular design of DISCO allows users to leverage a flexible combination of features according to their requirements. Such as offering choices on federated and decentralized communication protocols, various levels of privacy guarantees and several approaches to model learning that allows model personalization and resilience to unseen bias.

Please contact either of the authors to find out more.

**References**

1. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 2021 37. 2021;3: 610–619. doi:10.1038/s42256-021-00338-7

2. Schwarz CG, Kremers WK, Therneau TM, Sharp RR, Gunter JL, Vemuri P, et al. Identification of Anonymous MRI Research Participants with Face-Recognition Software. N Engl J Med. 2019;381: 1684–1686. doi:10.1056/nejmc1908881

3. Hutson M. Taking the pain out of data sharing. Nature. 2022;610: 220–221. doi:10.1038/D41586-022-03133-5

4. 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. doi:10.1016/J.JCLINEPI.2022.05.019

5. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. npj Digit Med 2020 31. 2020;3: 1–7. doi:10.1038/s41746-020-00323-1

6. Roschewitz D, Hartley M-A, Corinzia L, Jaggi M. IFedAvg: Interpretable Data-Interoperability for Federated Learning. 2021 [cited 1 Nov 2021]. Available: https://arxiv.org/abs/2107.06580v1

**Authors and Affiliations**

**Intelligent Global Health Research Group, EPFL, Lausanne, Switzerland**

Mary-Anne Hartley, [mary-anne.hartley@epfl.ch](mailto:mary-anne.hartley@epfl.ch)

**Pediatric Emergency Unit, University Hospital of Bern, Inselspital, Bern, Switzerland**

Mary-Anne Hartley

**Machine Learning and Optimization Laboratory, EPFL, Lausanne, Switzerland**

Martin Jaggi, [martin.jaggi@epfl.ch](mailto:martin.jaggi@epfl.ch)

**Corresponding author**

Correspondence can be addressed to [mary-anne.hartley@epfl.ch](mailto:mary-anne.hartley@epfl.ch)

**Contributions**

MAH wrote and conceptualized the manuscript. MJ reviewed and contributed to all versions. MAH and MJ collaborate on developing the DISCO platform, originally conceptualized by MJ.

**Ethics Declarations**

**Competing Interests**

The Authors declare no Competing Financial or Non-Financial Interests

**Acknowledgments**

The authors would like to thank the dedicated team of students and engineers developing DISCO