Philos. Technol. (2018) 31:473–479
https://doi.org/10.1007/s13347-018-0320-8

COMMENTARY

# Democratizing Health Research Through Data Cooperatives

**Alessandro Blasimme**[1] · **Effy Vayena**[1] · **Ernst Hafen**[2]

Received: 30 May 2018 / Accepted: 31 May 2018 / Published online: 19 June 2018


**Abstract** Massive amounts of data are collected and stored on a routine basis in virtually all domains of human activities. Such data are potentially useful to biomedicine. Yet, access to data for research purposes is hindered by the fact that different kinds of individual-patient data reside in disparate, unlinked silos. We propose that data cooperatives can promote much needed data aggregation and consequently accelerate research and its clinical translation. Data cooperatives enable direct control over personal data, as well as more democratic governance of data pools. This model can realize a specific kind of data economy whereby citizens and communities are empowered to steer data use according to their motivations, preferences, and concerns. Policy makers can promote this model by recognizing citizens' rights to access and to obtain a copy of their own data, and by funding distributed data infrastructures piloting new data aggregation models.

**Keywords** Data cooperatives · Data sharing · Big data · Precision medicine · Ethics · Governance

Innovative data mining capabilities, such as natural language processing and machine learning, can detect clinically relevant patterns in an ever-expanding sea of health data. Emerging paradigms like precision medicine (Collins and Varmus 2015; Blasimme and Vayena 2017) and digital health (Vayena et al. 2018) are premised on such advanced capabilities (Hawgood et al. 2015). However, before deploying such tools, phenotypic, clinical, lifestyle, and multi-omic data—including data generated directly by patients and healthy individuals—must become available for analysis. This is not happening at the desired pace, as different kinds of individual-patient data reside in disparate, unlinked silos (Tenopir et al. 2011; Blasimme et al. 2018).

✉ Ernst Hafen
hafen@imsb.biol.ethz.ch

[1] Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland

[2] Department of Biology, ETH Zurich, Institute of Molecular Systems Biology, Zurich, Switzerland

We propose that data cooperatives can promote much needed data aggregation and consequently accelerate research and its clinical translation. The rationale for adopting data cooperatives is that people (healthy and sick) are the legitimate controllers of their personal data. Data cooperatives offer tools for exerting direct control over personal data, and for participating in the democratic governance of data pools. This model can realize a specific kind of data economy whereby citizens and communities are empowered to pull multifarious types of data in one place, and steer data use according to their motivations, preferences, and concerns. Policy makers can promote the creation of data cooperatives by recognizing citizens' right to access and to obtain a copy of their own data, and by funding the creation of distributed data infrastructures piloting new data aggregation models.

## 1 Translational Impediments

Large personal data repositories are being built in many countries (Ginsburg and Phillips 2018). The US Precision Medicine Initiative is assembling data from one million Americans. In parallel, the Department of Veterans Affairs is creating another large-scale research cohort through the Million Veteran Program. Genomics England is collecting DNA and clinical records from 100,000 UK citizens to develop precision medicine for rare diseases and cancer. Switzerland has launched its national Swiss Personalized Health Network in 2016 to create a nationwide data-exchange infrastructure. In the private sector, US-based healthcare provider Kaiser Permanente is putting together a research biobank collecting samples and data from half a million people. Despite the expectations underlying these initiatives, many obstacles stand in the way. Different countries show different levels of preparedness in terms of both appropriate data governance mechanisms and adequate data infrastructures (OECD 2017). Complex issues regarding informed consent, data portability, and privacy protections for research conducted on biomedical big data exemplify regulatory hurdles to progress in data-driven research (Blasimme et al. 2018). Poor interoperability, lack of data curation, and insufficient or poorly representative case control cohorts may foreclose research and clinical translation in the years to come (Blasimme et al. 2018). The chief impediment in this area, however, lies precisely at the intersection of governance and operational readiness. Massive amounts of data are collected and stored on a routine basis in virtually all domains of human activities. Such data are potentially useful to biomedicine. Patients and healthy citizens show increased inclination to make their data available for research purposes (Oliver et al. 2011). However, data reside in silos, as potent disincentives stand in the way of data sharing: scientific competition, unwillingness to grant access to existing datasets, regulatory burdens associated with releasing data to third parties, to name but a few. Publicly sponsored repositories of reference data (such as NIH's Genomic Data Commons) have compensated for those tendencies. Yet, incentives for aggregating individual data at the scale needed to enable data mining techniques have so far been limited. In parallel, the private sector is accumulating gigantic amounts of data under exclusive control, which is giving rise to concrete risks of data monopolies (Wilbanks and Topol 2016).

Novel data governance models can offer effective incentives for data sharing. Recently, there has been an upsurge of interest for governance schemes that recognize a greater role to

research participants (O'Doherty et al. 2011). The creation of large-scale biobanks in Europe as well as in North America has spurred intense debate about more accountable, participatory, and empowering forms of engagement. For instance, the reuse of personal information initially collected for a given study is now only possible if participants consent to it (Blasimme et al. 2017). New informed consent models and regulatory concepts such as data portability give data subjects increased control on the ever-growing amount of data they provide for research (Vayena and Blasimme 2017). This progressive recognition of participants' entitlements is gaining traction, with the All Of Us program spearheading the transition from research participation to a partnership-based model of research (NIH; Blasimme and Vayena 2016). While these are laudable developments, individuals still have little if any possibility to take a truly active role in data collection, governance, and distribution, as data are still sparkled in a multitude of largely unlinked sources. Data cooperatives, we argue, can tackle this bottleneck.

## 2 Data Cooperatives

The driving idea of data cooperatives is that individuals are the most legitimate actors to promote personal data aggregation and to claim data control (Wilbanks and Topol 2016). As such, the individual becomes the connecting node of her personal data scattered in disparate collections. A data cooperative member can be any individual who stores health-relevant data of virtually any type and format, such as data imported from mobile apps, electronic health records, and clinical data, as well as multi-omic data generated by research studies, on the cooperative platform (Fig. 1). In this way, members act as aggregators of their own data from multiple separated sources. Research groups can request access to these data. Oversight mechanisms within the cooperative ensure scientific and ethical assessment of incoming data access requests, but data subjects keep control over whether they want to grant access to their data and under which conditions. Consent is managed electronically directly through the cooperative. Members also take part in the governance of the platform, either directly or indirectly (for instance, electing the members of oversight committees), and exert collective control of the whole dataset, decide how revenues will be reinvested, and craft policies for specific activities or issues (such as the handling of incidental findings). In this way, cooperatives offer their members concrete opportunities for engaging in the democratic governance of stored data.

The MIDATA cooperative (www.midata.coop) jointly created in 2015 by ETH Zurich (Computer Science Department and Institute of Molecular Systems Biology) and the Bern University of Applied Sciences (Institute for Medical Informatics) is an example of a not-for-profit data cooperative that is already enabling access to data for research purposes. This platform is based on transparent governance principles and state-of-the-art encryption to ensure privacy. Moreover, it has a modular architecture and governance structure, meaning that sister cooperatives can be constituted in other regions and countries. Given that—in contrast to healthcare systems—the needs of citizens and patients are similar in different countries, studies (e.g., patient-reported outcomes) developed on the MIDATA platform in one country can easily be replicated in other countries.

Evidence shows that, while people are aware of privacy risks in sharing sensitive data, they give more weight to the benefit of sharing in privacy-utility determinations (Oliver et al. 2011). People's positive attitude towards sharing data—even highly sensitive ones

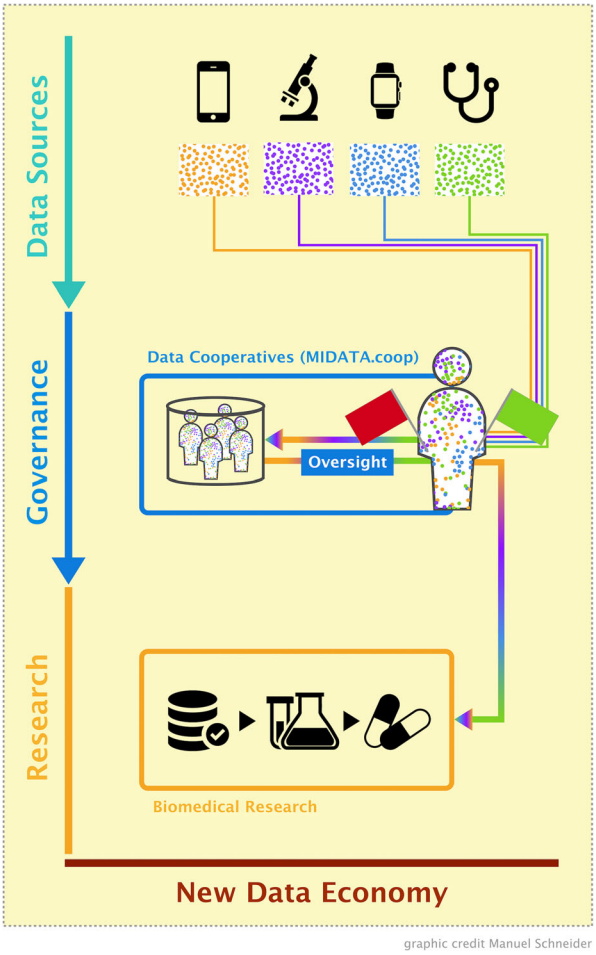


**Fig. 1 A new data economy for health research**. Citizens aggregate data from different sources and make them available for research through data cooperatives. Cooperatives offer oversight mechanisms to filter data access requests and tools for the democratic governance of the data

like genetic data—can promote data aggregation and distribution. Data cooperatives thus capitalize on people's willingness to make their data available for research studies that are perceived as important for a given community, a given disease or for society at large. Given that data subjects act also as data controllers, cooperative platforms are better placed to ensure the level of trust that is needed in data-driven health research.

With biomedical research venturing into the era of big data, like other enterprises in the digital economy, it faces critical challenges, such as the risk of data monopolies and lack of adequate reward for user-contributed assets. Data cooperatives, in this respect, share the motives of other movements calling for more democratic practices in the digital economy (Frenken 2017). Data cooperatives offer an innovative template to foster the direct engagement of data subjects in biomedical research, thus redressing the power asymmetry between them and data controllers—especially those in the private sector (Wilbanks and Topol 2016).

## 3 Expected Benefits and Challenges

As members of a data cooperative, individuals acquire the new role of data aggregators—and thus become an enabling factor of research. In contrast to conventional models of research participation, in which participants are minimally, if at all, involved in data governance processes, data cooperatives' members govern the data platform.

Data cooperatives, however, do not serve individuals' interests alone. They also cater to the needs of communities as they can be the interface for disease constituencies and other stakeholders to engage with data-driven health research. Less empowered or historically underserved communities can organize themselves and their specific projects on cooperatives allowing them to take control of their data and to voice their motivations, needs, and worries. This possibility can help redress the underrepresentation of minorities in health research databases, thus limiting the scientific and clinical consequences of using biased datasets (Landry et al. 2018). Given the growth of nationwide data repositories, cooperatives can act as tools to foster inclusion of otherwise marginalized social groups, allowing them to have their voices heard and to influence the direction of scientific activities (Sabatello 2017).

Cooperatives can offer scientists access to aggregated data that simply did not exist in this linked format before, thus enabling new data mining capabilities relying both on existing and on newly generated data. A scenario of this type allows research funders to free up resources from the de novo creation of large-scale data repositories. Importantly, cooperatives like MIDATA incentivize data curation by reimbursing hospitals that provide health information in standardized formats (e.g., LOINC, SNOMED, and FHIR)—on the assumption that such data can generate revenue to be re-invested in the cooperative itself. In turn, partnership between data cooperatives and researchers engages both parties to adopt appropriate data standards.

Data cooperatives' members can also allow their clinicians to access their data. In a not-too-distant future, this can have transformative effects allowing better monitoring and management of health and disease trajectories through real-world data (Real-World Evidence Generation and Evaluation of Therapeutics 2017). The industry can contribute to this transformation by meeting the demand for data analysis and interpretation platforms that can have direct application in healthcare practice. Regulatory agencies will have to provide adequate guidance in this domain. Given the rapid penetration of smartphones with its included sensors, dedicated apps provided by hospitals after treatment, which the patients download on their smartphone, store the data in their data cooperative account and agree that anonymized data to be shared with the hospital will provide real-world patient-reported outcome data on treatments and medications. The patients thus take an active role in improving evidence-based medicine.

Potential challenges for this governance model include funding and the public's commitment. To ensure financial sustainability in the absence of corporate investment, anonymized data authorized by the data subjects can be made available for a fee to private-sector parties under non-exclusive license agreements. However, the not-for-profit status of data cooperatives ensures that data can only be monetized to fulfill statutory aims—hence, in agreement with members' expectations.

While patients may anticipate direct benefit from data sharing, healthy people may have less incentive to claim copies of their data and to join a cooperative. Still,

as people become increasingly aware of the risks of data monopolies, freeing one's data from inaccessible silos, making them available for the public good of science and participating in their governance, configures a new data economy model that will appeal to many—independently of their health conditions.

## 4 Policy Needs

The creation of data cooperatives can be stimulated through ad hoc funding strategies. For instance, in the context of flagship scientific initiatives, research funders can promote data cooperatives through calls that emphasize pilot research projects through distributed data aggregation platforms. In this way, data cooperatives, possibly alongside other models, will channel streams of well-curated individual data into scientific activities and ensure a sustainable development for self-governed data repositories.

Moreover, legislation should be in place to ensure that citizens have a right to access their data and to obtain a copy of them. This will facilitate data aggregation and enable data subjects to use their data for a variety of other purposes as well (e.g., seeking second opinions, or setting up a citizen science project). Promising changes are underway. Data access and data portability rights—recently sanctioned by the European General Data Protection Regulation—are a good illustration of this tendency. In the USA, the Blue Button initiative allows patients to download their health records (Turvey et al. 2014). Moreover, a public consultation launched by the Office of Science and Technology Policy under the Obama Administration in late 2016 showed interest in further increasing data portability in the USA (Exploring Data Portability 2016). Data portability implies both that data subjects can claim a copy of their personal data from data controllers, and that they can have their data transferred to another controller. This right enables data subjects to promote data aggregation and to take up roles conventionally not available to research participants—including those of citizen scientists and collective stakeholders.

A longstanding issue in need of policy response is data ownership. Legal understandings of data ownership vary from jurisdiction to jurisdiction. Data protection laws clearly identify data controllers, yet ownership is ill defined. Therefore, the distinction between these two concepts remains elusive, adding uncertainty to data initiatives and undermining the emergence of alternative data sharing platforms. In order to promote data availability, clear guidance should be in place to elucidate if and how the prerogatives of data controllers relate to personal data ownership.

Facilitating the development of data cooperatives through focused funding, legislative incentives, and regulatory clarification offers novel solutions to the bottleneck of data sharing. Furthermore, the scope of data cooperatives can be adapted to different contexts in different countries and thus promote the democratization of the personal data economy and of biomedical research globally. In this way, data cooperatives can become powerful gateways to individual health data, enabling researchers to access richer datasets in the context of a direct, fiduciary relation with data subjects.

## References

Blasimme, A., & Vayena, E. (2016). Becoming partners, retaining autonomy: ethical considerations on the development of precision medicine. *BMC Medical Ethics, 17*(1), 67. https://doi.org/10.1186/s12910-016-0149-6.

Blasimme, A., & Vayena, E. (2017). "Tailored-to-you": public engagement and the political legitimation of precision medicine. *Perspectives in Biology and Medicine, 59*(2), 172–188.

Blasimme, A., Moret, C., Hurst, S. A., & Vayena, E. (2017). Informed consent and the disclosure of clinical results to research participants. *The American Journal of Bioethics, 17*(7), 58–60. https://doi.org/10.1080/15265161.2017.1328532.

Blasimme, A., Fadda, M., Schneider, M., & Vayena, E. (2018). Data sharing for precision medicine: policy lessons and future directions. *Health Affairs, 37*(5).

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *The New England Journal of Medicine, 372*. https://doi.org/10.1056/NEJMp1500523.

Exploring Data Portability (2016). Office of science and technology options.

Frenken, K. (2017). Political economies and environmental futures for the sharing economy. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 375*(2095). https://doi.org/10.1098/rsta.2016.0367.

Ginsburg, G. S., & Phillips, K. A. (2018). Precision medicine: from science to value. *Health Affairs, 37*(5), 694–701.

Hawgood, S., Hook-Barnard, I. G., O'Brien, T. C., & Yamamoto, K. R. (2015). Precision medicine: beyond the inflection point. *Science Translational Medicine, 7*(300), 300ps317. https://doi.org/10.1126/scitranslmed.aaa9970.

Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs, 37*(5), 780–785.

NIH About the All of Us Research Program. https://allofus.nih.gov/about/about-all-us-research-program.

O'Doherty, K. C., Burgess, M. M., Edwards, K., Gallagher, R. P., Hawkins, A. K., Kaye, J., et al. (2011). From consent to institutions: designing adaptive governance for genomic biobanks. *Social Science & Medicine, 73*(3), 367–374. https://doi.org/10.1016/j.socscimed.2011.05.046.

OECD. (2017). *New health technologies*. Paris: Organisation for Economic Co-operation and Development.

Oliver, J. M., Slashinski, M. J., Wang, T., Kelly, P. A., Hilsenbeck, S. G., & McGuire, A. L. (2011). Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Public Health Genomics, 15*(2), 106–114. https://doi.org/10.1159/000334718.

Real-World Evidence Generation and Evaluation of Therapeutics (2017). The National Academies of Sciences, Engineering, and Medicine.

Sabatello, M. (2017). Precision medicine, health disparities, and ethics: the case for disability inclusion. *Genetics in Medicine*. https://doi.org/10.1038/gim.2017.120.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: practices and perceptions. *PLoS One, 6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101.

Turvey, C., Klein, D., Fix, G., Hogan, T. P., Woods, S., Simon, S. R., et al. (2014). Blue button use by patients to access and share health record information using the Department of Veterans Affairs' online patient portal. *Journal of the American Medical Informatics Association, 21*(4), 657–663. https://doi.org/10.1136/amiajnl-2014-002723.

Vayena, E., & Blasimme, A. (2017). Biomedical big data: new models of control over access, use and governance. *Journal of Bioethical Inquiry, 14*(4).

Vayena, E., Haeusermann, T., Adjekum, A., & Blasimme, A. (2018). Digital health: meeting the ethical and policy challenges. *Swiss Medical Weekly, 148*, w14571.

Wilbanks, J. T., & Topol, E. J. (2016). Stop the privatization of health data. *Nature, 535*(7612), 345–348. https://doi.org/10.1038/535345a.