Unlocking the potential of patient data through responsible sharing – has anyone seen my keys?

Patrick Cullinan[1], **Liz Roberts**[2]

[1] bluebird bio, Inc, Cambridge, MA, USA
[2] UCB Biosciences, Inc, Raleigh, NC, USA

Correspondence to:

Patrick Cullinan
bluebird bio
60 Binney Street,
Cambridge, MA 02142 USA
pcullinan@bluebirdbio.com

Abstract

The sharing of individual participant-level clinical data is now an almost routine extension of the clinical study life-cycle, and increasingly a vital element of leveraging real-word data. Responsible clinical study data sharing of appropriately consented and de-identified participant-level data and associated clinical documents is an expectation of key research funders, journal editors, pharmaceutical trade associations, regulators, ethics committees, and government entities sponsoring research. Furthermore, patients increasingly support expanded data sharing to help spur innovation and maximise the utilisation of data gathered during clinical studies. Finally, rapidly and appropriately leveraging real-world data to support and validate clinical research data and to facilitate responses to emerging public health emergencies lends greater importance and urgency to finding better ways to unlock and share health data. This article provides an overview of the current state of participant-level data sharing in clinical research and a discussion of the opportunities that exist to better navigate barriers to access whilst respecting the data privacy rights of study participants. This article describes our collective journey through the data sharing ecosystem, looking to further unlock the value of study participant data to drive new discoveries.

Background

Data obtained through clinical research are fundamental to advancing the field of medicine and to improving the health and well-being of patients. The data underlying such research have historically remained securely in the custody of the data generator (in most cases a pharmaceutical or academic study sponsor) and access was highly limited. Recently there have been increasing calls from patient groups, advocacy organisations, journal editors, pharmaceutical and biotechnology trade associations, regulators, and others in the scientific community for the responsible sharing of study patient-level data-sets and/or study documentation, to provide greater transparency and propel research innovation through secondary data reuse. Additionally there have been calls by many, including the World Health Organization (WHO), to accelerate and extend these data sharing paradigms to speed up the collection and dissemination of data during public health emergencies, a need exemplified during the recent Ebola outbreaks in West Africa and the COVID-19 pandemic.[1] It is envisaged that by harnessing the statistical power of large data-sets the broader scientific community can embrace the "big data" revolution, including utilising machine learning and artificial intelligence to spur new frontiers in data analytics and data collaboration.[2] Recently, all data generators have been further incentivised to share data by journal requirements to make their data sharing plans public as a pre-requisite commitment for publications (such as those aligned to the International Committee of Medical Journal Editors position and associated with PLoS One journals).[3,4] In addition, in 2016 the FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for Scientific Data Management and Stewardship[5] were published to provide guidelines to data generators, to improve the findability, accessibility, interoperability, and reuse of data which has further helped to propel data sharing, especially from the academic perspective. As such, all clinical research data generators have experienced increasing requirements and calls to establish mechanisms to responsibly share the clinical study data they produce.

This article provides an overview of patient-level data sharing with a focus on clinical trial data sharing, a discussion of key limiters and enablers that require greater attention to optimally unlock the value of patient-level data, and aspects of data sharing that have yet to be fully harnessed to further drive new discoveries.

Data sharing process

Patient-level data sharing refers to the process whereby data providers accept requests from academic and non-academic researchers for access to data and supporting documentation from formal clinical trials. Figure 1 provides an overview of the common steps of the data sharing request process that generally occur, although the process may vary based on the data provider or access platform.

Discoverability

Before a requester can make a data sharing request they need to know what data are available and understand key data characteristics. As such, there is a need for data providers to make public their data sharing policies and to communicate the availability of studies.

With this in mind, the trade association principles for responsible clinical study data sharing require sponsors to make public their data sharing policies and provide a mechanism to receive Research Proposals (RPs). In addition, ClinicalTrials.gov recently implemented an IDP (individual patient data) sharing plan section of the register that must be completed by each sponsor as part of study registration to improve the discoverability of data. As such, most pharmaceutical and biotechnology sponsors have made their data sharing policies and processes public, and increasingly data generators of all types are making their data sharing plans public at the time of study registration and publication. In most cases, pharmaceutical data providers have joined consortia or created stand-alone portals that specifically list studies for sharing to aid discoverability. Together, these measures have

dramatically increased the discoverability of clinical studies available for data sharing RPs and are helping to spur a new era of transparency and data-driven innovation.

Applying for access

Data access requires submission of a robust RP by a requester (usually a researcher on behalf of a broader research team that includes a statistician or suitably qualified data-analytics professional [e.g., health economist]). The RP requests specific studies and notes other data that the researchers will seek to aggregate or otherwise include in their analysis. In many cases, RPs also include data management and statistical analysis plans that outline precisely how they will manage and use the requested data, as well as detailed publication plans and conflicts of interest statements.

Review, approval, contracting, and access

In most cases, RPs are initially reviewed by the data provider for completeness and alignment to an organisation's data sharing policies, the study informed consent, and other legal bases for sharing (e.g., consistent with the European Union [EU] General Data Protection Regulation [GDPR]).[6] Requests are subsequently reviewed by an independent review panel (IRP) to assess the scientific merit and other aspects of the request (e.g., conflicts of interest and researcher qualifications). The manner in which IRPs are utilised (e.g., as the primary review panel or as an appeal panel for sponsor-rejected requests), their role in review, and the degree of independence varies. Upon RP approval researchers and/or their institution must sign a data sharing agreement (DSA) specifying the data access conditions and licences that are being granted. These agreements include a commitment to protect the privacy of study participants and the confidentiality of data provider information, and detail other obligations and rights associated with data access.

Once the DSA is executed, access to anonymised data and de-identified documents are provided, in most cases, via a secure cloud-based research environment. Data protections seek to minimise the risk of participant/patient re-identification and release of company confidential information. The protected data are generally provided for a defined period of time (usually 1 to 2 years), although extensions are possible.

Data sharing landscape

The rapid evolution of data sharing has resulted in the development of a complex and often non-interoperable landscape of data sharing platforms and research environments. This overview provides a high-level landscape summary of the types of data sharing systems and a summary of key platforms that have developed, although it is not intended to represent an exhaustive list. Rather, it is intended to provide a sense of the types of platforms that have developed. For specific information, policies, and processes relating to a specific sharing mechanism, the reader should refer to the applicable data sharing portals or provider websites.

Pharmaceutical study sponsor data sharing

Although ad hoc and fit-for-purpose data sharing has been occurring for some time in the pharmaceutical industry, large-scale and coordinated data sharing implementation in a broader sense gathered momentum in response to the establishment of trade association principles for responsible data sharing in 2014.[7] Early adopter companies developed mechanisms to share data through the establishment of portals to accept requests, IRPs to adjudicate access, and secure data sharing research environments. These early efforts to share data led to the creation of ClinicalStudyDataRequest.com (CSDR) and The YODA Project (Yale Open Data Access) portals.[8,9] Other pharmaceutical entities have created similar partnerships to facilitate data sharing, for example, Duke Clinical Research Institute has partnered with pharmaceutical sponsors and others to facilitate access via Duke's Supporting Open Access for Researchers (SOAR) platform.[10] While there are distinguishing differences between these portals (for example, some aspects of SOAR emphasise curation and data harmonisation), they essentially follow the

same major steps outlined in Figure 1 (study listing, proposal review, IRP approval, DSA, research conduct). One limitation of these platforms is that there has been little interoperability of the research environments established for these entities, although some recent efforts have been made to permit researchers to request data across these and other systems.

While these consortia/academic-supported data sharing platforms have made rapid progress, the majority of pharmaceutical and biotech sponsors share data via stand-alone portals (ranging in complexity from proposal submission gateways to more simple online forms or email request systems) and utilising various company-specific IRPs and data access approaches.

Another pharmaceutical sponsor data sharing arena is related to pre-competitive data sharing intended to permit collaboration amongst sponsors to spur more efficient and effective clinical development. Examples of such pre-competitive data sharing include the IQ (Innovation and Quality) consortium of pharmaceutical and biotechnology companies who share (largely) technically-focused pre-clinical and early clinical data to identify new science, technology, and regulatory engagement pathways.[11] Another is the DataCelerate platform developed to support sharing amongst TransCelerate and BioCelerate member companies.[12]

While the extent of growth in the area of data access has been rapid and impressive, the proliferation of platforms has resulted in data access mechanisms that have limited interoperability and are inefficient and difficult to navigate from a researcher perspective. Efforts are underway to create greater opportunity for cross-platform access and improve the efficiency of the process overall.

Non-profit data sharing portals

While pharmaceutical sponsors have been expanding efforts to directly share data with each other and with independent researchers, non-profit entities have been entering the data sharing space, both in their capacity as funders and through the creation of data sharing infrastructure to further facilitate access, lower the threshold for entry, and imparting further independence to the process. An important non-profit active in this space is the Wellcome Trust, which has been a leader in developing and enforcing data sharing requirements for its funded research and has also served to support the development of both CSDR and Vivli (see below), primarily by supporting the management of IRPs on these platforms.[13]

Direct efforts by non-profit organisations (often associated with specific funders or patient organisations) to support data sharing are largely

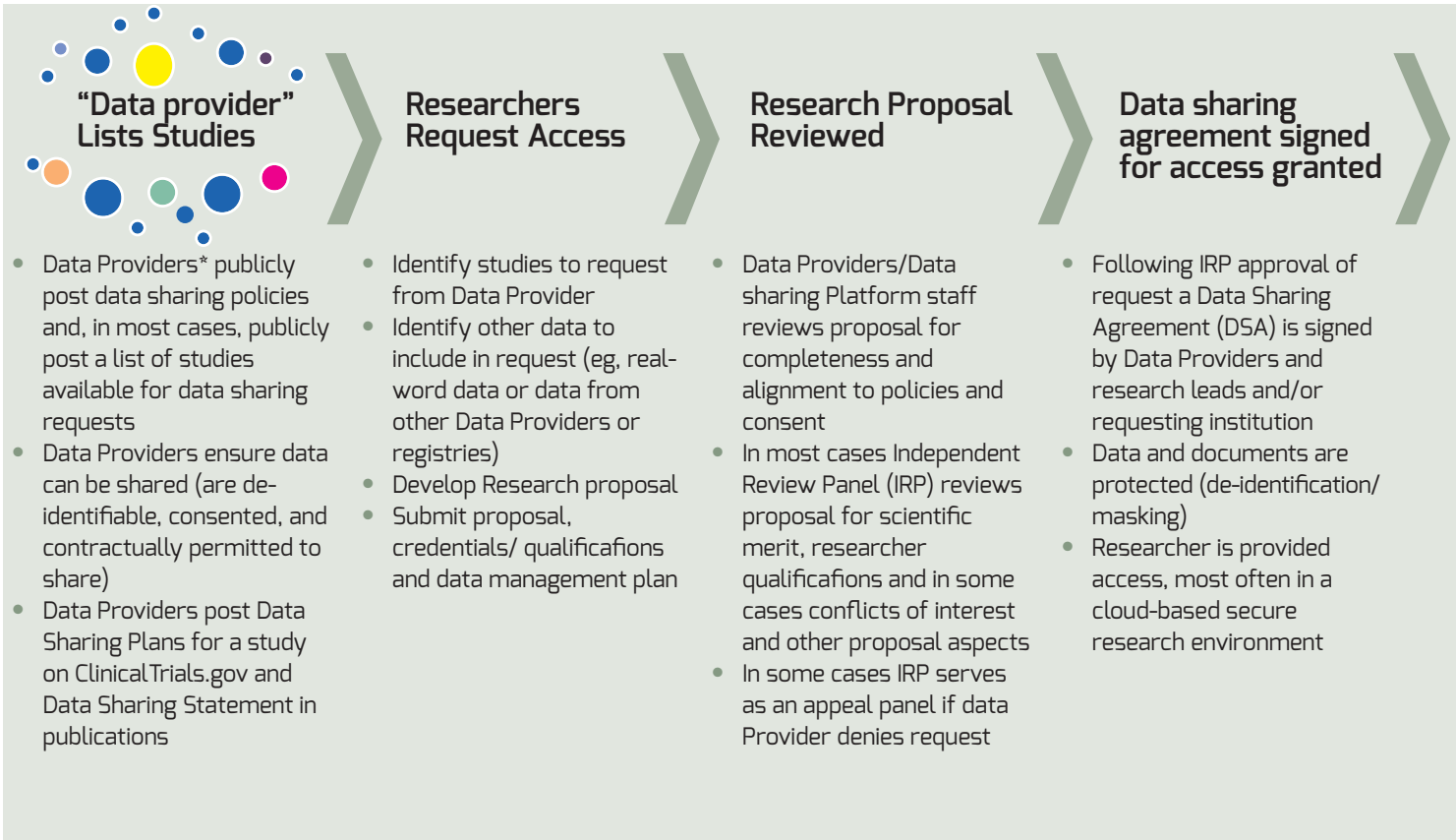

Figure 1. Common data sharing request steps

Not all data providers or data sharing mechanisms require all steps or in some cases may include extra steps (such as an appeal process for denied access requests).

*Data Provider is a broad term intended to encompass all data generators including both academic and pharmaceutical/biotech sponsors and other data generators (e.g., non-profit research entities). It also includes data requested from data stewards who manage acquired data that they did not generate (such as through an acquisition, merger, or via a data donation).

disease- or therapeutic-area specific. An important early example of such sharing is the oncology-specific Project Data Sphere, which was created by *CEO Roundtable on Cancer's Life Sciences Consortium* and which initially focused on the sharing of downloadable comparator-arm data from oncology studies, although this platform also has the capacity to host data in a secure environment or more recently via Vivli (see below).[14]

Another important non-profit data sharing entity is Vivli, which is a data sharing platform that seeks to serve as a neutral broker between data providers, data accessors, and the wider data sharing community (that includes pharmaceutical, academic, non-profit, and public-private data).[15] Vivli has developed a global data sharing and analytics platform that seeks to span all disease areas, facilitate interoperable data sharing across a range of data providers, and intends to expand to create disease-specific communities and add capability to support data protection processing (de-identification) to lower the barrier to entry for smaller entities and academic data providers.

Public-private partnerships

The efficiency and impact of data sharing can be maximised when focused curation and harmonisation of data-sets occurs. As such, data sharing is increasingly an important part of large public-private partnerships (PPPs), which are large initiatives coordinated by governmental (public) and industry/academic (private) entities and created through shared public and private funds. These partnerships generally focus on specific public health priorities and, in doing so, can bring substantial resources and organisational infrastructure to accelerate innovation through enhanced data sharing and other methods.

In Europe, the Innovative Medicines Initiative (IMI) is a partnership between the EU and the European Federation of Pharmaceutical Industries and Associations to address a range of important healthcare research topics. More and more of IMI's projects include data sharing/aggregation aspects focused on enabling access and innovative new research relating to specific diseases.[16] In the US, similar PPPs are funded by the Foundation for the National Institutes of Health with the support of the Pharmaceutical Research and Manufacturers of America and are called Accelerating Medicines Partnerships (AMPs).[17] AMPs seek to accelerate research in a range of disease areas and to support biomarker development. Another PPP model in the US is coordinated by the Critical Path Institute which is a non-profit PPP involving the FDA that aims create new data, measurement, and methods standards through the aggregation of data to spur new innovation in a pre-competitive consortium model via their platforms.[18]

Other data sharing

Patient-level clinical study data sharing is increasingly supplemented by data from other sources. Data from patient registries and patient data aggregation projects (e.g., the UK biobank and the US National Institute of Health "All of Us" campaign), as well as real-world data from electronic health records and other sources such as wearable devices, will increasingly be leveraged by researchers to supplement or compare against clinical study data.[19, 20] Importantly, while the ability to leverage such sources may improve the power of data sharing, such data have limitations in terms of its quality, uniformity, and standardisation. In addition, combining such data sources with clinical study data may represent an increased risk to participant re-identification as the possibility of linking and identifying patients across larger and more diverse data-sets increases.

> By harnessing the statistical power of large data-sets the broader scientific community can embrace the "big data" revolution, including utilising machine learning and artificial intelligence to spur new frontiers in data analytics and data collaboration.

Data yet to be fully unlocked

While there have been substantial advances in the field of data sharing, as outlined below, there are several types of data that are not optimally being shared.

Rapid data sharing during public health emergencies

For the most part, the data sharing mechanisms outlined previously take a deliberate approach to data sharing that seek to responsibly account for the privacy and consent of participants and protect the intellectual rights of researchers and sponsors to protect confidential information and data rights. Such "responsible data sharing" therefore takes a methodical approach that can be potentially time consuming and it may not be possible to share certain data due to confidentiality, privacy, and other legal limitations. Such sharing is not suitable for rapid data dissemination as is needed during a public health emergency. Insufficient timely access to reliable data severely hampers epidemiological tracking to the spread of disease and efforts to coordinate control and implement treatment responses and research. Our collective deficiencies in rapidly collecting and sharing basic scientific data (such as viral gene sequences), real-world data (such as sharing infection rates, patient symptoms, disease

Research conducted and project closed

- Researcher conducts project per proposal analysis plan
- Researcher request extension if needed or amends project (with appropriate approvals)
- Researcher notifies Data Providers/IRP of any new safety signal, intellectual property or publications
- Access to the project is closed per DSA
- Manuscript submitted for publication

trajectory, epidemiological data, and outcomes to treatment protocols), and early clinical trial data in a manner that is responsible, timely, accurate, and interoperable, to appropriately inform public health policy was identified during the recent Ebola outbreaks in West Africa in 2014 to 2016, the subsequent emergence of Zika infections, and more recently the COVID-19 pandemic.[21,22] Unfortunately with every outbreak there is an urgent need to establish or re-establish ad hoc mechanisms for expedited data collection, sharing, and publication rather than implementing the utilisation of established standards and systems. An example of an effort to provide a mechanism for such sharing is the recent implementation by the WHO of a "COVID-19 Open" data sharing and reporting protocol, which seeks to provide a mechanism for rapid online publishing of COVID-19 research papers – similar to approaches implemented following the emergence of Zika infections.[1,21,22] These rapid peer-reviewed publications are among the many efforts to share results that the WHO and others have repeatedly attempted to implement during health emergencies, yet clearly there is a need to proactively establish infrastructure and data standards for the collection and sharing of data during health emergencies that can overcome geopolitical, language, and other barriers and support informed scientific research and health-policy decisions through more timely sharing of standardised data.

To this end, at the time of writing, certain pharmaceutical companies, not-for-profit organisations, academia, and health authorities have united across various platforms to explore new ways to collaborate and responsibly share data more promptly. How this sustainably changes the paradigm of data sharing post-pandemic is yet to be seen.

Rare-disease data

Sharing of rare-disease data represents an innovation opportunity if challenges relating to patient privacy can be overcome. Aggregation of rare-disease patient-level data can help overcome the paucity of patients participating in research, for example, by providing historical control data. However, most data providers have policies that exclude sharing any data where the risk of re-identification of patients would be elevated, and as such, do not share data from studies in diseases considered rare. To overcome this issue, the broader scientific community is working to develop advanced data anonymisation and sharing technology (possibly utilising encryption, synthetic data modelled on the actual data, or employing distributed analytic techniques that bring the analytics to the data [rather than sharing the data itself]) and enhanced patient consents that more clearly consent patients by outlining the risks of re-identification and potentially offering patients the option to opt out of sharing, and alternative legal bases for sharing and managing rare-disease patient data.

Genomic data and biospecimen sharing

Genomic data by their very nature are unique to a given individual and so represent immense potential that is limited by privacy concerns. In addition, these data are very sensitive, and their misuse could have implications beyond an individual (e.g., having implications for family members and a potential generational impact). As such, efforts to broadly share genomic data and to tie the sharing of such data to clinical study data or other data sources have been limited, although early examples have emerged (e.g., the Psychiatric Genomics Consortium) and mechanisms are being developed to better enable such sharing.[23] Similarly, biospecimen/sample sharing represents another underutilised data resource that has been limited by concerns related to consent, import/export and privacy regulations, re-identification risk, logistical matters, and lack of clarity related to "ownership" and data-steward responsibility for newly-generated data from the sample. Enabling better utilisation of genomic and biospecimen data in a responsible manner with adequately informed and consented patients and leveraging new technologies to protect patient privacy will be important to unlocking the huge potential of these data sources.

Keys to drive enhanced data sharing

While there has been substantial recent progress towards enhanced FAIR access to participant-level data, there remain substantial barriers that continue to limit access and hamper the efficiency of the ecosystem and medical communication professionals can play an important role in unlocking the data.

One important way in which medical communication professionals can enable data sharing is to consider this topic in the development of informed consents and protocols. Clearly discussing data sharing in these documents can facilitate later data sharing. Informing patients of the sponsor's data sharing plans, possible use of such data, and residual privacy risks associated with sharing of de-identified data can substantially enable future data sharing. Such consent and protocol language can help clarify the legal basis of sharing as it relates to evolving privacy legislation and streamline ethics committee approval.

Medical communicators also play an important role in improving discoverability through including appropriate and clear data sharing plans on clinical trial registers and publications. Furthermore, medical writers can support subsequent data sharing processes by employing lean-writing approaches that minimise the need to redact while still producing documents of high clinical utility. Indeed, making data and documents easier to protect can enhance utility to the research community and can make the sharing process more efficient.

Other important enablers of data sharing efficiency include broadening the use of common data standards and more prospectively releasing information about the structure and contents (data dictionary/metadata) of "to be shared" data. Posting such information along with the listed title and basic metadata would allow researchers to more effectively plan and understand "what they are getting", thus enabling more efficient and successful data sharing.

Conclusions

Data sharing has made immense progress in the past five years yet more can be done to unlock its true potential, especially considering emerging disease challenges that will require data driven solutions. Medical writers are well positioned to be a key contributor to facilitate progress in this space. Making data more discoverable, improving protocol and patient consent language relating to data sharing and the associated residual risks, ensuring clear description of the legal basis for sharing, and improving the timeliness, efficiency, and utility of shared data and documents through lean authoring and writing with privacy protection in mind, can substantially unlock and enable enhanced data sharing.

Disclaimers

The opinions expressed in this article are the authors' own and not necessarily shared by their employers or EMWA.

Conflicts of interest

The authors are employed by bluebird bio (and a bluebird stock owner) and UCB Biosciences, respectively, and either in the context of their current or past employers have supported or participated in some of the data sharing platforms discussed in this paper, particularly ClinicalStudyDataRequest.com and Vivli.org.

References

1. Modjarrad K, Moorthy VS, Millett P, Gsell PS, Roth C, Kieny MP. Developing global norms for sharing data and results during public health emergencies. PLoS Med. 2016 Jan;13(1):e1001935.
2. Institute of Medicine. Sharing clinical trial data: Maximizing benefits, minimizing risk. Washington DC: The National Academies Press; 2015.
3. Taichman DB, Backus J, Baethge C, et al. A disclosure form for work submitted to medical journals: a proposal from the International Committee of Medical Journal Editors. Lancet. 2020 Jan;S0140-6736(20)30187-2.
4. Bloom T. PLOS' New Data Policy: Part Two. In: EveryONE: PLOS ONE Community Blog [internet]. 2014 [cited 2020 Mar 06]. Available from: https://blogs.plos.org/everyone/2014/03/08/plos-new-data-policy-public-access-data/
5. Wilkinson M, Dumontier M, Aalbersberg I, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016:160018. doi: 10.1038/sdata.2016.18.
6. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation [GDPR]).
7. European Federation of Pharmaceutical Industries and Associations, Pharmaceutical Research and Manufacturers of America. Principles for Responsible Clinical Trial Data Sharing: Our Commitment to Patients and Researchers. 18 Jul 2013 [cited 2020 Mar 06]. Available from: https://efpia.eu/media/25189/principles-for-responsible-clinical-trial-data sharing.pdf
8. Clinical Study Data Request Consortium data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://www.clinicalstudydatarequest.com/
9. Yale Open data Access data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://yoda.yale.edu/
10. Duke Supporting Open Access for Researchers (SOAR) data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://dcri.org/our-work/analytics-and-data-science/data sharing/
11. International Consortium for Innovation and Quality in Pharmaceutical Development (IQ) data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://iqconsortium.org/about/
12. DataCelerate data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://transceleratebiopharmainc.com/datacelerate/
13. Wellcome Trust Data, software and materials management and sharing policy [internet]. [cited 2020 Mar 06]. https://wellcome.ac.uk/grant-funding/guidance/data-software-materials-management-and-sharing-policy
14. Project Data Sphere data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://projectdatasphere.org/projectdatasphere/html/home
15. Vivli data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://vivli.org/
16. the Innovative Medicines Initiative Projects and Results [internet]. [cited 2020 Mar 06]. Available from https://www.imi.europa.eu/projects-results
17. Foundation for the National Institute of Health Accelerating Medicines Partnerships [internet]. [cited 2020 Mar 06]. Available from: https://fnih.org/what-we-do
18. Critical Path Institute data access portals [internet]. [cited 2020 Mar 06]. Available from: https://c-path.org/programs/
19. UK Biobank data sharing platform [internet]. [cited 2020 Mar 06]. Available from: https://www.ukbiobank.ac.uk/
20. All of US data access platform [internet]. [cited 2020 Mar 06]. Available from: https://www.researchallofus.org/workbench/
21. Dye C, Bartolomeos K, Moorthy V, Kieny MP. Data sharing in public health emergencies: a call to researchers. Bull World Health Organ. 2016 Mar;94(3):158.
22. Moorthy V, Henao Restrepo AM, Preziosi M-P, Swaminathan S. Data sharing for novel coronavirus (COVID-19). Bull World Health Organ. 2020 Mar;98(3):150.
23. Sullivan PF, Agrawal A, Bulik CM, et al. Psychiatric genomics: An update and an agenda. Am J Psychiatry. 2018;175(1):15–27.

Author information

Patrick Cullinan, PhD, is Senior Director of Medical Writing for bluebird bio, a gene therapy company in Cambridge Massachusetts. In addition to medical writing, Patrick has broad experience leading clinical trial transparency and data sharing teams, and was a former member of ClinicalStudyData Request.com and Vivlii.org steering committees and has supported data sharing planning and initiatives for a range of non-profit and public-private partnerships organisations.

Liz Roberts, MSc., is an experienced Senior Director with almost 30 years' experience within pharmaceutical R&D. Transitioning from a career in biostatistics and with a demonstrated history in the field of clinical data transparency, disclosure, and patient-level data sharing, she provides external thought-leadership to increase the value these activities can bring to people living with severe diseases. Particular areas of interest include value-based transparency, the legal framework supporting data sharing activities, and the co-creation of forward-looking, meaningful policies, and deliverables.