Integrated but Isolated: Implications from a Systematic Review of the Access Control Ecosystem for Individual Participant Data in Clinical Studies

Lee, Jian-Sin University of Michigan, USA | jianslee@umich.edu
Jeng, Wei National Taiwan University, Taiwan | wjeng@ntu.edu.tw

ABSTRACT

While the importance of open science is further highlighted during the pandemic, the challenges of managing and sharing individual participant data (IPD) derived from clinical studies never cease. The nature of IPD, e.g., confidentiality or sensitivity, makes it difficult to maintain a good balance between data sharing and individual privacy protection. To date, many access control mechanisms for IPD do exist, but conventional solutions and services are deemed scattered and still not in place. To gain a more comprehensive understanding of the IPD sharing tensions, we conducted a systematic literature review with 64 academic publications that discuss the access control mechanisms built for IPD in clinical studies. Via the knowledge infrastructure (KI) framework, we identified nine key aspects involved and the relationships between major stakeholders in the IPD access control ecosystem. Our results anticipate informing the future design of an IPD management checklist that data professionals can use to guide their clients when releasing sensitive biomedical data.

KEYWORDS

Individual participant data; Clinical data; Access control; Data sharing; Data management plans

INTRODUCTION

A clinical study refers to research recruiting human volunteers (i.e., participants) in which the research team learns if a specific medical intervention is more effective and/or has less harmful side effects than the conventional one. Clinical studies aim to find ways to prevent disease, improve quality of life (e.g., smoking cessation studies), or enrich medical knowledge (U.S. NLM, n.d.). A clinical team is charged by a principal investigator (often a medical doctor) and comprises other doctors, nurses, and healthcare professionals. Clinical studies conducted by the team can be sponsored by government agencies, academic institutions, or pharmaceutical companies.

Due to the heavy workload, sites that conduct clinical trials currently tend to resort to technologies such as clinical trial management systems (CTMS) or health information systems (HIS) to manage the huge amount of clinical data generated during different clinical phases (Park et al., 2018), as well as coordinate the staffing. Among the overarching clinical data, e.g., study descriptions for clinical trial registration as administrative metadata and clinical study reports as summary data (Institute of Medicine, 2015), *individual participant data* (IPD) serve as a critical source for data reuse because they include fine-grained raw data and analyzable datasets, such as medical images and case report forms of certain trial participants. This type of data is especially essential for IPD meta-analysis (IPDMA), a “gold standard” approach in evidence-based medicine that combines IPD from multiple studies into a single dataset for further analysis (Riley et al., 2010; Stewart & Clarke, 1995).

Calls for transparency of clinical studies have been raised since the late 1990s through the 2000s worldwide: regulatory authorities and journals thus started requiring trial registration (DeAngelis et al., 2005; U.S. FDA, 1997). A trial registration is the practice of registering the basic information about a clinical trial before it is conducted. Following that, as major funding agencies announced their data sharing mandates in the 2000s, the importance of sharing general clinical data has gradually come to light. However, it was not until the 2010s that IPD sharing has become a focused open science issue, due to their confidentiality and sensitivity.

IPD sharing is considered an ethical obligation because of human participants’ sacrifices for public interests (Taichman et al., 2017). The National Institutes of Health (NIH) (2018) released its Strategic Plan for Data Science, where one of the objectives is to incorporate high-value clinical resources, including IPD, into its biomedical data science system. In this plan, the NIH simultaneously restated that biomedical research data should stick to the FAIR Principles, emphasizing data being *findable*, *accessible*, *interoperable*, and *reusable*. Nonetheless, to protect human subjects’ rights, the International Committee of Medical Journal Editors (ICMJE) (n.d.) required manuscripts submitted to its affiliated journals to include a data sharing statement that specifies the appropriate mechanisms by which data would be accessed. The NIH (2022) will also renew its data management plan (DMP) template in 2023 by adding more elements pertaining to data access, reuse considerations, and oversight of data sharing.

Owing to the sensitive nature of IPD, data holders often grant limited access to those who intend to reuse IPD to strike a balance between data openness and trial participants’ privacy (Hopkins et al., 2016; Sydes et al., 2015). The reconciliation between integration- and isolation-based data sharing strategies, however, is not easy. Tensions have

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

thus been observed in terms of stakeholders' practices of IPD sharing and data protection. Many IPD access mechanisms have been built in multiple forms and at various levels of control, e.g., centralized data platforms or emerging technologies such as encryption-related blockchain techniques (Benchoufi & Ravaud, 2017). Nonetheless, despite some explorations of these IPD access control mechanisms, existing tools and services are deemed scattered, therefore hard to discover and perform quality assessments (Ohmann et al., 2018). Such challenges have posed a pressing need for a more holistic understanding of these pivotal methods to release IPD in a secure manner.

This preliminary study thus seeks to synthesize 64 scholarly publications through a systematic literature review (SLR) to 1) identify the key aspects involved when examining IPD access control mechanisms and 2) reveal the relationships between major stakeholders by mapping the IPD access control ecosystem. As we view data sharing as a knowledge sharing practice, we apply the concept of knowledge infrastructure (KI) as a guidance framework to break down and able to depict the complex and interwoven ecosystem of IPD sharing. KI was defined as "robust networks that generate, share, and maintain specific knowledge about the human and natural worlds" (Edwards, 2010, as cited in Borgman, 2015, p. 33), of which we consider its seven elements a good fit for reflecting the dynamic and static IPD flow: people, artifacts, built technologies, institutions, policies, shared norms and values, and routines and practices (Edward et al., 2013, as cited in Jeng & He, 2022).

Based on the SLR results, our ultimate goal is to develop an *IPD management checklist* containing comprehensive modules and items that enable data stewards to produce thorough descriptions when drafting a DMP or an IPD sharing statement for their clients from medical fields. Our findings can also provide insights for researchers and pharmaceutical companies when deciding to release IPD and for developers when designing ideal mechanisms that make IPD sharing safer and easier. Since biomedical researchers demand professionally curated data and heavily rely on rigorous data management solutions (Borda et al., 2020; Wang et al., 2019), information science (IS) professionals working across disciplines are undoubtedly worth paying more attention to all such data-driven issues.

METHODS

To address our research aim at profiling the IPD access control ecosystem, an SLR approach is considered suitable because it can be helpful for exploring the knowledge frontiers and discovering potential gaps (Xiao & Watson, 2019). Our sampling process consists of a two-phase literature search. The first phase (covering 52 publications) ended in May 2021, and the second phase, which followed the same search protocol, was stopped in April 2022 to accommodate and catch up with more current literature. Our final sample includes 64 articles.

We first conducted the literature search in three databases: Web of Science, Scopus, and PubMed. Google Scholar was then selected as the fourth source of literature to ensure comprehensiveness. We used the following query string to search the fields of publication title, abstract, and keywords: *("clinical studies" OR "clinical trials") AND ("individual participant data" OR "individual patient data") AND "data sharing"*. During the search process, we did not apply terms such as "access control" because 1) a pilot search showed that adding these terms generated limited results, and 2) discussions on IPD access control mechanisms were often embedded in the general literature regarding IPD sharing.

Our initial literature list contains 826 publications. Referring to the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) statement, after identifying potentially relevant publications (n=423) and then removing the duplicates (n=151), we carried out an eligibility check on 272 publications. To meet the inclusion criteria, 1) a healthy proportion (i.e., at least one full paragraph) of a publication should be dedicated to access control mechanisms for IPD, 2) the publication should be a journal or conference article, a technical report, or a book chapter written in English, and 3) full text of the publication should be available. A total of 64 publications were eventually included in our SLR. On the Open Science Framework (OSF) platform, we provide a few materials as an OSF project (https://osf.io/5vjm4/) for references: the list of analyzed literature, the full literature search protocol, queries used for each database, detailed visualizations of the respective sampling process for the first-/second-phase and overall searches, and descriptive characteristics of the sampled literature.

RESULTS

Most of the literature we reviewed (L01-L64, see List S1 in our OSF project) were published after 2012 (n=62, 96.9%), showing an upward trend in numbers by year (Figure S1). Nearly half of the sample (n=28) was authored by U.S. research teams and the other 31 by European teams, e.g., the UK (n=14) and Germany (n=7). The context where the literature was published matters because data protection regulations can vary significantly by country or region and thus be essential for future in-depth examinations.

Applying the concept of KI, a strong theoretical framework that helps to establish attributes for infrastructures in data sharing research (Jeng & He, 2022), we identified nine key aspects of IPD access control mechanisms. In Figure 1, we present four major components in the access control ecosystem for IPD, namely the clinical trial team (T), its sponsors (S), the data requesters (R) intending to access IPD, and the access control mechanisms (A). The

IPD per se and each of the nine aspects corresponds to one of the seven KI elements and can be used to describe the relationships among the four components (Table 1). Next, we briefly illustrate the review results of each aspect.

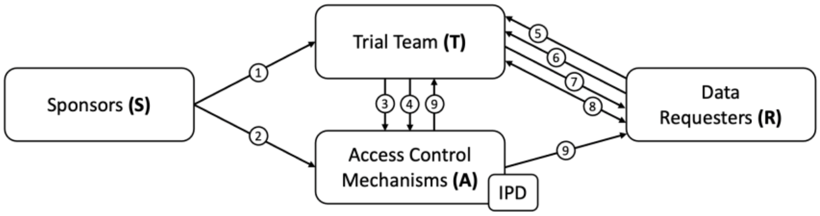

Figure 1. The Access Control Ecosystem for IPD

Aspect	KI Element	Description
-	artifacts	IPD in clinical studies
funder requirements	policies	1. S applies its funding requirements to T.
time frames	routines & practices	2. S specifies the timelines when IPD must be made available via A.
technological solutions	built technologies	3. T selects appropriate technical solutions as A.
request methods	routines & practices	4. T establishes IPD request process for A.
eligibility	people	5. R satisfies the conditions set by T.
purposes	shared norms & values	6. R reports its intent to access IPD to T.
review committees	institutions	7. T organizes a panel to review data requests from R.
agreements/statements	policies	8. Documents that manage IPD use are drafted between T and R.
costs	routines & practices	9. A necessitates time and resource investments from T and R.

Table 1. Descriptions of Each Aspect and the Corresponding KI Element

Note. S, T, A, and R respectively stand for sponsors, the trial team, access control mechanisms, and data requesters.

Funder requirements. As the controller of IPD, clinical trial sponsors take legal responsibility for ensuring participants' privacy (L36). Different sponsors' IPD sharing requirements for trial teams could therefore vary. When compared with nonindustry ones, industry sponsors were found to be more active in making clinical data sharing policies as well as express willingness to share IPD and provide information about the availability of various types of data upon registration, probably because academic teams have fewer resources and thus are less motivated to make such efforts (L12, L13). Besides, industry sponsors also share IPD via data platforms more often, which is considered relatively efficient to coordinate with multiple parties (L02, L55).

Time frames. Embargoes and duration of use are two common measures implemented for IPD access control in terms of timelines. A one- or two-year embargo is usually laid to secure the original team's intellectual property by allowing them to perform their own secondary analysis (L09, L49). In some cases, the trial data are made available once certain stages in the clinical trial lifecycle are completed, e.g., after the resulting paper is published (L01, L39). As for the duration of use, IPD access is often granted for one to three years, which can be extended if necessary (L24, L32). However, it is worth noting that studies show a low proportion (20-30 percent) of both funders' policies and trialists' registration records having properly specified the time frames for IPD sharing (L13, L56).

Technological solutions. As a critical component in the ecosystem, technological solutions are adopted by IPD holders to implement access control. In particular, data platforms usually create data enclaves, where requesters can analyze (but cannot download) the data in a secure (e.g., password-protected) environment (L02, L26). Granular access to different units of data also helps to manage the extent a single dataset is shared (L40, L42). More alternatives include algorithms and querying languages that facilitate data use without accessing them (L03, L28), as well as Bitcoin- and blockchain-based techniques that can encrypt and track data for security (L10).

Request methods. Many data platforms require IPD requesters to undergo a step-by-step process before data access is granted, e.g., the Clinical Study Data Request (CSDR) platform (L10, L62). Typical steps include: 1) requesters submit a research proposal, 2) an independent committee reviews the application, 3) if approved, the study sponsor grants permission, 4) requesters sign a data use agreement, 5) requesters access IPD and conduct analysis in a protected environment, and 6) requesters report any resulting publications to the data platforms. Other options, such as the Project Data Sphere (PDS) platform (L15, L45), maintain a simpler process that allows less scientific review and thus are considered more "open."

Eligibility. "Who the data requesters are" determines whether they are eligible to access IPD. Some data repositories prefer minimal control over who can access IPD by only requiring user sign-up, which is a more open-access practice (L14). There are even calls for the right to access IPD for any citizens and groups with reasonable needs (L39). Other data platforms following stricter procedures tend to consider IPD requesters' expertises, qualifications, and whether significant conflicts of interest exist (L12). A special case is the European Medicines Agency (EMA) policy on publication and access to clinical trial data, of which an older version stated that a data requester should be established in the EU (L01).

Purposes. Levels of access control also depend on the expected purposes of data requesters. IPD holders might be reluctant to share data given the concerns about data dredging or reanalysis with unknown intent (L30). Although some open clinical data repositories might encourage data sharing without asking about requesters' purposes, access to controlled data is often granted for enhancing public health (L01, L14). Secondary analysis and IPDMA are the most frequently mentioned purposes for accessing IPD (L05, L38). Other possibilities include exploratory analysis, reproducibility tests, replication, verification, and developing statistical or trial methods (L08, L12, L47).

Review committees. During the data request process, a review committee usually serves as a gatekeeper deciding whether the access should be granted. Besides a few internal committees, most of the committees are external to ensure independence (L01, L24). External committees can be further distinguished by whether the members were selected and paid by the data holder itself or a third party. Concerning the committee members, cross-disciplinary experts are often preferred, e.g., physicians, statisticians, and epidemiologists. Representatives from medical ethicists, patient advocates, and sponsors are frequently discussed as well (L12, L57).

Agreements or statements. Written documents such as a data use agreement (DUA) or data sharing statement (DSS) are frequently used to provide details about how IPD access will be controlled. For example, to mitigate the risks to trial participants, IPD holders draft DUAs to explicitly define data requesters' rights, obligations, and liabilities for IPD use (L16, L49), e.g., the analyses allowed and the responsibility for maintaining confidentiality (L37). With a DSS, opportunities are also provided for trial teams to specify their willingness to share data and the data access methods when registering trials or submitting their resulting manuscripts to journals (L09, L31).

Costs. Dealing with IPD access control is considered expensive, time-consuming, and labor-intensive for both IPD requesters and holders (L21, L43). Specifically, IPD requesters can take a long time to receive data after submitting a request, depending on data holders. It is reported that obtaining IPD from repositories takes the least amount of time (less than one month), followed by original researchers (approximately half a year) and centralized data platforms (four months to a year) (L32, L50, L62). For IPD holders, operations that generate costs include establishing a platform with a request system, responding to data requests, organizing review panels, negotiating DUAs with requesters, and so on (L24).

DISCUSSION & CONCLUDING REMARKS

Our preliminary study findings depict the picture of the IPD access control ecosystem by using the knowledge infrastructure (KI) framework to identify its nine key aspects and the connections between major stakeholders. Our approach is bottom-up, and the early results can be conducive to establishing the building blocks of an IPD management checklist. Readers can thus make good use of it and introduce accurate and practical guidance to their clients for preparing to release sensitive biomedical data, such as IPD in clinical studies. As more and more IPD access control mechanisms are in place, our review results can help them facilitate scattered data sources to be integrated efficiently, and at the same time be securely stored and well-protected in their home institutions.

By revisiting the goals of KI set over the past decade, Borgman and colleagues (2020) recently pointed out the increasing fragility and brittleness of KI when facing the era of open data. Funding agencies have also recognized the difficulties in guaranteeing continued access to the ever-growing amount of biomedical data (Borgman, 2020). Such challenges in sustaining critical biomedical data infrastructures further highlight the potential of the access control ecosystem we portrayed and of the to-be-derived checklist to inform the data management and sharing requirements from funding agencies or journals. For example, the NIH's current DMP template (2021) has hardly addressed issues regarding data request methods, restrictions imposed by funders, and roles overseeing the data sharing process, each of which can be mapped to one of the nine aspects we identified. The same or a tailored combination of these building blocks is believed to enable similar examinations into requirements or guidelines such as the NIH's forthcoming new DMP template in 2023 (U.S. NIH, 2022), the ICMJE's DSS template (Taichman et al., 2017), and other clinical trials toolkits (e.g., Clinical Development Services Agency, n.d.).

Moving forward, to develop our checklist, gathering empirical data from interviews or surveys is necessary to further understand how data access control mechanisms work in the real world and how trial teams' perspectives influence their practices on these mechanisms. It is also worth deep-diving into some of the emerging topics from the reviewed literature, e.g., patient involvement in the decision-making process for IPD sharing and the reduced data utility after IPD are de-identified. All in all, to empower the IPD sharing enterprise, we recommend that IS professionals take action by keeping a close watch on relevant trends in the latest policies and standards.

ACKNOWLEDGMENTS

This work was financially supported by the Ministry of Science and Technology (MOST) in Taiwan, under MOST 111-2636-H-002-004- and MOST 111-2634-F-002-018-, the Center for Research in Econometric Theory and Applications (Grant no. 111L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project, and also the Universities and Colleges Humanities and Social Sciences Benchmarking Project (Grant no. 111L9A002) by the Ministry of Education (MOE) in Taiwan. This work was done while Jian-Sin Lee was at the Department of Library and Information Science, National Taiwan University, Taiwan. Please address all correspondence to Wei Jeng (wjeng@ntu.edu.tw).

REFERENCES

Benchoufi, M., & Ravaud, P. (2017). Blockchain technology for improving clinical research quality. *Trials, 18*: 335.

Borda, A., Gray, K., & Fu, Y. (2020). Research data management in health and biomedical citizen science: Practices and prospects. *JAMIA Open, 3*(1), 113-125.

Borgman, C.L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Borgman, C. L. (2020). Knowledge infrastructures in past, present, and future tense. *UCLA: Center for Knowledge Infrastructures*. Retrieved April 17, 2022, from https://escholarship.org/uc/item/5v73333z

Borgman, C. L., Darch, P. T., Pasquetto, I. V., & Wofford, M. F. (2020). Our knowledge of knowledge infrastructures: Lessons learned and future directions. *UCLA: Center for Knowledge Infrastructures*. Retrieved April 17, 2022, from https://escholarship.org/uc/item/9rm6b7d4

Clinical Development Services Agency. (n.d.). Clinical Trials Toolkit – India: Data Management Plan. Retrieved April 17, 2022, from https://cdsatoolkit.thsti.in/data-management-plan/

DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., ... & Van Der Weyden, M. B. (2005). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *Archives of Dermatology, 141*(1), 76-77.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges*. Ann Arbor: Deep Blue. http://hdl.handle.net/2027.42/97552

Evangelatos, N., Reumann, M., Lehrach, H., & Brand, A. (2016). Clinical trial data as public goods: Fair trade and the virtual knowledge bank as a solution to the free rider problem – A framework for the promotion of innovation by facilitation of clinical trial data sharing among biopharmaceutical companies in the era of omics and big data. *Public Health Genomics, 19*(4), 211-219.

Hopkins, C., Sydes, M., Murray, G., Woolfall, K., Clarke, M., Williamson, P., & Smith, C. T. (2016). UK publicly funded Clinical Trials Units supported a controlled access approach to share individual participant data but highlighted concerns. *Journal of Clinical Epidemiology, 70*, 17-25.

Institute of Medicine. (2015). The clinical trial lifecycle and when to share data. In *Sharing clinical trial data: Maximizing benefits, minimizing risk*. Washington, DC: The National Academies Press.

International Committee of Medical Journal Editors. (n.d.). Clinical Trials. Retrieved April 17, 2022, from https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

Jeng, W., & He, D. (2022). Surveying research data-sharing practices in US social sciences: A knowledge infrastructure-inspired conceptual framework. *Online Information Review, ahead-of-print*. doi: 10.1108/OIR-03-2020-0079

National Institutes of Health. (2018). NIH Strategic Plan for Data Science. Retrieved April 17, 2022, from https://datascience.nih.gov/nih-strategic-plan-data-science

National Institutes of Health. (2021). NIH-GEN: Generic (Current until 2023). Retrieved April 17, 2022, from https://dmptool.org/template_export/1888.pdf

National Institutes of Health. (2022). NIH-GEN DMSP (Forthcoming 2023). Retrieved April 17, 2022, from https://dmptool.org/template_export/118304408.pdf

Ohmann, C., Canham, S., Banzi, R., Kuchinke, W., & Battaglia, S. (2018). Classification of processes involved in sharing individual participant data from clinical trials. *F1000Research, 7*: 138.

Park, Y. R., Yoon, Y. J., Koo, H., Yoo, S., Choi, C. M., Beck, S. H., & Kim, T. W. (2018). Utilization of a clinical trial management system for the whole clinical trial process as an integrated database: System development. *Journal of Medical Internet Research, 20*(4): e9312.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ, 340*: c221.

Stewart, L. A., & Clarke, M. J. (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine, 14*(19), 2057-2079.

Sydes, M. R., Johnson, A. L., Meredith, S. K., Rauchenberger, M., South, A., & Parmar, M. K. (2015). Sharing data from clinical trials: The rationale for a controlled access approach. *Trials, 16*: 104.

Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., ... & Backus, J. (2017). Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. *The New England Journal of Medicine, 376*, 2277-2279.

U.S. Food and Drug Administration. (1997). Food and Drug Administration Modernization Act of 1997. Retrieved April 17, 2022, from https://www.govinfo.gov/content/pkg/PLAW-105publ115/pdf/PLAW-105publ115.pdf

U.S. National Library of Medicine (n.d.). Learn About Clinical Studies. Retrieved April 17, 2022, from https://clinicaltrials.gov/ct2/about-studies/learn

Wang, X., Williams, C., Liu, Z. H., & Croghan, J. (2019). Big data management challenges in health research: A literature review. *Briefings in Bioinformatics, 20*(1), 156-167.

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research, 39*(1), 93-112.