Title

A hierarchical strategy to minimize privacy risk when linking De-identified data in biomedical research consortia.

Authors

Ohno-Machado, Lucila

Jiang, Xiaoqian

Tao, Shiqiang

et al.

Publication Date

2023-03-01

DOI

10.1016/j.jbi.2023.104322

Peer reviewed

HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2024 March 28.

Published in final edited form as:

J Biomed Inform. 2023 March ; 139: 104322. doi:10.1016/j.jbi.2023.104322.

A hierarchical strategy to minimize privacy risk when linking "De-identified" data in biomedical research consortia

Lucila Ohno-Machado[a,c,1,*], **Xiaoqian Jiang**[b,1], **Tsung-Ting Kuo**[a], **Shiqiang Tao**[b], **Luyao Chen**[b], **Pritham M. Ram**[b], **Guo-Qiang Zhang**[b], **Hua Xu**[b,c]

[a]UCSD Health Department of Biomedical Informatics, University of California San Diego Health, La Jolla, CA, USA

[b]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

[c]Biomedical Informatics & Data Science, Yale School of Medicine, New Haven, CT

Abstract

Linking data across studies offers an opportunity to enrich data sets and provide a stronger basis for data-driven models for biomedical discovery and/or prognostication. Several techniques to link records have been proposed, and some have been implemented across data repositories holding molecular and clinical data. Not all these techniques guarantee appropriate privacy protection; there are trade-offs between (a) simple strategies that can be associated with data that will be linked and shared with any party and (b) more complex strategies that preserve the privacy of individuals across parties. We propose an intermediary, practical strategy to support linkage in studies that share de-identified data with Data Coordinating Centers. This technology can be extended to link data across multiple data hubs to support privacy preserving record linkage, considering data coordination centers and their awardees, which can be extended to a hierarchy of entities (e.g., awardees, data coordination centers, data hubs, etc.)

1. Introduction

Given the rapid increase in data available for biomedical research, [1,2] the potential to link data from the same participant across different Studies, as well as to link data from other types of datasets (e.g., Electronic Health Records - EHRs), is increasingly higher. More complete and linked data will yield a better picture of an individual's health over time and consequently provide better insights for biomedical discoveries. Unfortunately, linked data make it easier to identify individuals. [3,4] There has been a lot of confusion in the biomedical research community about the role of Privacy-Preserving Record Linkage

[*]Corresponding author: Lucila.Ohno-Machado@Yale.edu (L. Ohno-Machado).

[1]Co-first authors

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104322.

(PPRL) methods [5,6], which is aimed at finding whether the record of the same individual is present in different databases without revealing the individual's identity (also referred to as blind data linkage,[7] private data linkage,[8] etc. by different researchers). The primary purpose of PPRL is to protect the identities of individuals (especially those who are not matched in the process) by comparing two databases without having to reveal the identities of individuals. Table 1.

Many researchers have studied the PPRL problem and developed various solutions based on different privacy technology frameworks, including bloom filters [9,10], embedding space [11], generalization [12], scrambling [13], differential privacy [14], secure multiparty computation (SMC) [15,16], and homomorphic encryption [17]. These solutions offer different tradeoffs between privacy and utility, which reflects in the differences in speed/computation cost, security promise, need for a third-party broker, and the number of supported collaborators. These solutions are frequently applied in a flat, non-hierarchical fashion, leaving individuals more exposed to a potential breach.

Naive approaches using one-way hashing (without a secret value) are highly efficient and can mask personal data during record linkage, but they can still be subject to brute-force attacks such as dictionary attacks [18,19]. Dimension reduction techniques, like embedding space and generalization, also offer high efficiency, but they are vulnerable to frequency attacks [20]. On the other hand, perturbation techniques, like scrambling to introduce faked samples and differential privacy to change records slightly, offer better privacy protection at the cost of potential linkage errors. There is also a difference between the number of parties involved in the linkage process and their roles. The garbled circuit protocol can provide strong theoretical security guarantees to safeguard the linkage process of two parties but does not generalize well to more parties [21]. SMC approaches offer rigorous protections for-multiparty linkage, but they are computationally intensive (and do not scale up well because of the pair-wise comparison nature of encrypted linkage). All of these solutions are context-dependent, and no "one-size-fits-all" technique exists.

Other PPRL methods that use third party (honest) brokers have also been proposed.[22–24] This can be done by sending seeded hashed results to a third party, which will then match hashes from another data set and return the matched hash pairs to the original owner. Kho *et al.* developed a seeded hashing algorithm that uses predetermined fields to generate hashes to support record linkage with a semi-honest (curious but not malicious) third-party broker. [25] They tested this PPRL algorithm on 7 million records across 6 institutions in Chicago, and the algorithm was able to correctly match records with a high degree of accuracy. In a similar architecture, Datavant implemented commercial software to link records with a third-party broker [26], which is adopted by the National COVID Cohort Collaborative (N3C) and National Patient-Centered Clinical Research Network (PCORnet). Despite these advancements, data from different large health data initiatives are still relatively isolated, and there are not many solutions for a hierarchy of data collections to support efficient privacy preserving record linkage. The hierarchical solution mitigates the danger of successful dictionary attacks, since different hashed IDs are used at different levels, and it can be practical in the context of data coordinating centers.

We present a practical approach from the perspective of a Data Coordinating Center (DCC) [27] trying to collect and harmonize data from a large set of heterogeneous Studies, which generate data for the proposed research. We would like to ensure high efficiency multi-party record linkage with moderate privacy protection using a linkage unit that will only see encoded values. Data collected by DCC will then be sent, for example, to a National Institute of Health (NIH) Data Hub, which serves as the permanent repository for hosting and distributing the data among broader research communities. A simplified schematic of the Awardee-DCC-Data Hub relations is shown in Fig. 1. It should be noted that PHI data are only retained by the projects' researchers, and thus stay close to the source. Data are shared with the research community through the Data Hub only, which gathers data from multiple DCCs.

1.1. A privacy-preserving linkage strategy beyond typical PPRL methods

Our proposed method is as follows. First, every Study hashes (see Appendix A) participant identifiers into unique Study IDs (SIDs) (i.e. transforms identifiers such as name, date of birth, and/or social security number, into a token or string of characters and numbers that is unique to the participant, but cannot be reversed to reveal identifiers) that are submitted together with other data to the DCC.

However, using only SIDs is not enough. It may be possible to obtain participant consent for the linkage of data to other Studies after the data have been submitted to the DCC, or that a public health emergency requires the linkage even without authorization. Therefore, the second step is to generate Linkable IDs (LIDs), which share a common hashing seed. That is, every participant must have both a SID and an LID. A Study team will know the SID but not the LID (they are encrypted at source), which will only be known to the DCC. The LID is generated from the Study sites using the same identifiers as those used for the SID. The encrypted LID is sent to the DCC. Fig. 2 shows the IDs generation at the Awardee institutions and their transmission to the DCC.

Thus, the DCC gets the "de-identified" data plus the SID and the LID directly from the Study team, ensures harmonization and "de-identification", and submits the SID plus the other data to the Data Hub, which could be a data repository hosted by funding agencies such as NIH (National Institute of Health). The DCC has the SID and LID for a participant, but not any other identifiers. It does not need to retain the other data once the Data Hub has confirmed receipt and approval of the data. The DCC retains the SID-LID pairs to use when the Data Hub requests linkage across studies (Fig. 2B and 2C).

Finally, the DCC uses the LID when the linkage is requested to pair SIDs for a participant who has participated in various Studies and the DCC submits the LID-SID pair to the Data Hub. Through the entire process, the DCC and the Data Hub do not retain individual subject/participant identifiers. If linkage across DCCs is requested, the DCCs send their LIDs only to the Data Hub, which serves as an honest broker to support the permitted secure linkage for local institutions to connect patients through the LIDs. Once it determines which LIDs match, it sends the list to the DCCs, which then returns an ordered list of SIDs.

1.2. Investigator usage scenario for record linkage from several studies

If an investigator receives permission (the Data Access Committee or equivalent) to link the data of a group of participants in Group 1 studies, then the following are the steps for the Data Hub to link those participants' records (if they exist):

Step 1: Studies receive software (with preconfigured SID and LID seeds for hashing; the LID seed will be the same for all institutions and SID seeds are study-specific) to generate a pair of SID and LID based on patient identifiers. LIDs are not visible in the Studies.

Step 2: These pairs of SIDs and LIDs, as well as Study de-identified data about the participant, will be sent to DCC through an encrypted channel, and the Study will only retain non-linkable SIDs. Note that DCC never retains other types of identifiers on the server and instead just retains SID-LID pairs, so it can provide linkage when requested by the Data Hub.

Step 3: Given a data linkage request for Studies within the coordination of one DCC, the latter will check permissions and consent (indicated by SIDs provided by local institutions) to conduct linkage on LIDs. Note that only permitted comparisons between LIDs would occur to respect patients' privacy. LIDs will lead to the identification of the same patients across different Studies and the DCC will submit the final SID-LID pairs to the Data Hub. The final outputs of this process will allow linkable patients (under permission/consent) to connect SIDs while not touching any identifiable information.

Step 4: If the Data Hub requests linkage across Studies managed by different DCCs, then it will use the lists of LIDs from these DCCs to determine whether there are matches. It will submit the list of matched LIDs to the DCCs to obtain the SIDs. The Data Hub will only learn the pairing of SIDs and LIDs for matched participants. Here again, no identifiers will be available to the Data Hub other than the returned SIDs.

2. Scaling up PPRL to a hierarchy of entities (e.g., awardees, data coordination centers, data hubs, etc.)

The privacy protecting linkage of study data across DCCs can be extended to larger programs across different Data Hubs, such as those covering projects (in a consortium) at a national scale and across consortia managed through different research initiatives. We can generalize the concept of SID and LID to a set of different levels (e.g., LID1, LID2, LID3, etc.) to reflect the hierarchy of different entities. Fig. 3 schematically demonstrates our approach.

3. Discussion

There is a need for privacy preserving record linkage within and across health DCCs. This is because as health data are increasingly collected, linked, and used for research and other purposes, there is a greater risk of individuals' personal information being revealed. DCCs need to be able to link records while preserving the privacy of the individuals involved. It is important to have mechanisms in place that support PPRL considering a hierarchy of

DCCs and their awardees for better coordination with greater transparency. This allows each DCC to maintain and manage its own data while still being able to link records with others through data hubs and the “super” data hub in the hierarchy. Existing methods and solutions have not been designed with this scope in mind, and as such, they are not well-suited to address all the challenges that a DCC ecosystem presents.

Our proposed method is still in its conceptual stages, but it has the potential to reduce the risk of privacy breaches at Data Coordinating Centers (DCCs) and Data Hubs. Our solution is particularly noteworthy as it adheres to the principle of data minimization and supports the PPRL hierarchy of entities while keeping identifiers local. This has multiple benefits. For one, it minimizes the burden on the Study Data Coordinating Centers (DCCs), as they do not need to store and maintain any data other than simple SID-LID pairs. Additionally, it provides another layer of protection for participants in the studies, as the Data Hub only learns LIDs for matched participants. This ensures that the data is kept secure and private, and that participants’ identities are not revealed.

Neither DCC nor Data Hub identifiers, but there is still a possibility of linking their information in order for the Studies to recontact participants or to perform, if authorized, linkage across Studies. Evidently, if Studies match their data against a large portion of the linked data, they will learn additional information about their participants. For this reason, policy protections should also include provisions against retention of data at the Study level if there is not an intent to recontact the participant and no other linkages are anticipated. Once data has been accepted at the Data Hub, SIDS-LIDS pairs retained at the DCCs should be enough to allow proper linkages.

Our proposed solution is highly pragmatic, requiring only two rounds of seeded hashing between DCCs and the Data hub to generate SIDs and LIDs for privacy preserving record linkage. This solution is efficient and secure, and it can be implemented without any changes to the existing infrastructure. It naturally follows the governance structure of the DCCs and Data Hub, and is highly generalizable. The solution we proposed can be easily extended to tackle other data linkage tasks, including regional health surveillance and longitudinal studies for chronic diseases. With very few assumptions (e.g., no collusion between the linkage unit and participating sites), our proposed solution can support practical and high-efficiency record linkage with enhanced privacy protection in a multi-party scenario.

Our framework has a few limitations that should be taken into consideration. For example, the server is vulnerable to a denial-of-service attack, making it a single point of failure. Furthermore, the framework has limited ability to deal with missing data, which can reduce its effectiveness. Another limitation is that hashed-based record linkage models, when compared to other types of linkage methods, may have some difficulty determining which records should be linked. This can be especially challenging when two records have similar information but are not exact matches. As a result, hashed-based record linkage can be less accurate than other methods, such as probabilistic record linkage (e.g., techniques like bloom filters) that can handle misspelling and erroneous inputs (due to the different data quality from participating sites). However, a challenge of adapting probabilistic linkage is to support manual verification when multiple candidates appear

with similar matching scores. This might lead to an increase in attack exposure, as plain text data needs to be exposed to the linkage server, which may require additional patient consent or the development of alternative strategies to safeguard the verification process. Another probabilistic linkage strategy is to use privacy preserving federated learning to transform record linkage into a distributed classification problem. This would allow us to incorporate deep learning algorithms and convert patient records into embeddings to deal with misspellings and erroneous inputs while rendering probabilistic outputs. The gradient in the federated learning model can be encrypted before it is sent to the server, and then the server could perform the classification without decrypting the data, which protects the privacy of study participants in the probabilistic linkage process.

Finally, explaining the higher privacy risks when linking two seemingly "de-identified" data sets, even to a highly sophisticated scientific community, is not easy. Most biomedical researchers prefer to concentrate their efforts in their domain area, and are not aware of potential risks. Additionally, the fact that some linkage methods have "Privacy-Preserving" in their names can be misleading, as most refer to methods that have the primary ability to preserve the privacy of participants who were not matched, and less so the privacy of those whose data were effectively linked.

4. Conclusion

As more and more data get produced and made available for sharing, it is important that we keep developing strategies to minimize risk, cost, and adhere to one of the most important data principles, the one that refers to *Data minimization*, which implies "Adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."[28–30] This may increase trust in data sharing and further increase participation in research studies, particularly for those groups currently under-represented for a lack of trust in research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

The authors of this research are partially supported by the RADx-rad Data Coordination Center (DCC), which is funded by the National Institutes of Health (NIH) under grant U24LM013755.

References

[1]. Chen X, Gururaj AE, Ozyurt B, Liu R, Soysal E, Cohen T, Tiryaki F, Li Y, Zong N, Jiang M, Rogith D, Salimi M, Kim H-E, Rocca-Serra P, Gonzalez-Beltran A, Farcas C, Johnson T, Margolis R, Alter G, Sansone S-A, Fore IM, Ohno-Machado L, Grethe JS, Xu H, DataMed – an open source discovery index for finding biomedical datasets [Internet], Journal of the American Medical Informatics Association. (2018) 300–308, 10.1093/jamia/ocx121. [PubMed: 29346583]

[2]. Ohno-Machado L, Sansone S-A, Alter G, Fore I, Grethe J, Xu H, Gonzalez-Beltran A, Rocca-Serra P, Gururaj AE, Bell E, Soysal E, Zong N, Kim H-E, Finding useful data across multiple biomedical data repositories using DataMed, Nat Genet 49 (6) (2017 May 26) 816–819. [PubMed: 28546571]

[3]. Arellano AM, Dai W, Wang S, Jiang X, Ohno-Machado L, Privacy Policy and Technology in Biomedical Data Science, Annu Rev Biomed Data Sci 1 (2018 Jul) 115–129. [PubMed: 31058261]

[4]. Bonomi L, Huang Y, Ohno-Machado L, Privacy challenges and research opportunities for genomic data sharing, Nat Genet 52 (7) (2020 Jul) 646–654. [PubMed: 32601475]

[5]. Hall R, Fienberg SE. Privacy-Preserving Record Linkage. Privacy in Statistical Databases Springer Berlin Heidelberg; 2010. p. 269–283.

[6]. Clifton C, Kantarcioğlu M, Doan A, Schadow G, Vaidya J, Elmagarmid A, Suciu D. Privacy-preserving data integration and sharing. Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery New York, NY, USA: Association for Computing Machinery; 2004. p. 19–26.

[7]. Churches T, Christen P. Blind Data Linkage Using n-gram Similarity Comparisons. Advances in Knowledge Discovery and Data Mining Springer Berlin Heidelberg; 2004. p. 121–126.

[8]. Al-Lawati A, Lee D, McDaniel P. Blocking-aware private record linkage. Proceedings of the 2nd international workshop on Information quality in information systems ACM; 2005. p. 59–68.

[9]. Durham E, Kantarcioglu M, Xue Y, Toth C, Kuzu M, Malin B, Others. Composite Bloom filters for secure record linkage. IEEE Trans Knowl Data Eng IEEE; 2014;26(12):2956–2968.

[10]. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak BioMed Central Ltd; 2009;9(1):41.

[11]. Bonomi L, Xiong L, Chen R, Fung BCM, Frequent grams based embedding for privacy preserving record linkage, CIKM (2012) 1597–1601.

[12]. Karakasidis A, Verykios VS, Secure blocking + secure matching = secure record linkage, J Comput Sci Eng Korean Institute of Information Scientists and Engineers 5 (3) (2011 Sep 30) 223–235.

[13]. Karakasidis A, Verykios VS, Christen P. Fake injection strategies for private phonetic matching. Data Privacy Management and Autonomous Spontaneous Security Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 9–24.

[14]. Inan A, Kantarcioglu M, Ghinita G, Bertino E. Private record matching using differential privacy. Proceedings of the 13th International Conference on Extending Database Technology (EDBT) 2010. p. 123–134.

[15]. Lai P, Yiu S, Chow K, Chong C, Hui L. An Efficient Bloom Filter Based Solution for Multiparty Private Matching. Security and Management [Internet] 2006 [cited 2022 Aug 16]; Available from: https://www.semanticscholar.org/paper/ebbdf7b64eb8f8c8607b29c679db4b663a794d07.

[16]. Malin B, Airoldi E, Edoho-Eket S, Li Y. Configurable security protocols for multi-party data analysis with malicious participants. 21st International Conference on Data Engineering (ICDE'05) [Internet] IEEE; 2005. [doi: 10.1109/icde.2005.37].

[17]. Randall Brown, Ferrante. Privacy preserving record linkage using homomorphic encryption. First International Workshop on Population Informatics for Big Data (PopInfo'15) 2015.

[18]. Venot A, Burgun A, Quantin C. Medical Informatics, e-Health: Fundamentals and Applications. Springer Science & Business Media; 2013. ISBN:9782817804781.

[19]. Christen P, Ranbaduge T, Schnell R. Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing. Springer International Publishing; 2020.

[20]. Liu H, Wang H, Chen Y. Ensuring data storage security against frequency-based attacks in wireless networks. Distributed Computing in Sensor Systems Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 201–215.

[21]. OneFlorida Clinical Research Consortium [Internet]. 2015 [cited 2019 Jan 6]. Available from: http://www.pcori.org/research-results/2015/oneflorida-clinical-research-consortium.

[22]. Elfeky MG, Verykios VS, Elmagarmid AK. TAILOR: a record linkage toolbox. Proceedings 18th International Conference on Data Engineering ieeexplore.ieee.org; 2002. p. 17–28.

[23]. Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L, How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure, Int J Med Inform Elsevier 49 (1) (1998 Mar) 117–122.

[24]. Quantin C, Bouzelat H, Dusserre L. A computerized record hash coding and linkage procedure to warrant epidemiological follow-up data security. Stud Health Technol Inform 1997;43 Pt A:339–342. [PubMed: 10179568]

[25]. Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, Humphries JE, Kominers SD, Hota BN, Sims SA, Malin BA, French DD, Walunas TL, Meltzer DO, Kaleba EO, Jones RC, Galanter WL, Design and implementation of a privacy preserving electronic health record linkage tool in Chicago, J Am Med Inform Assoc The Oxford University Press 22 (5) (2015 Sep) 1072–1080.

[26]. Datavant: Connecting Health Data to Improve Patient Outcomes [Internet]. Datavant. 2021 [cited 2022 Nov 18]. Available from: https://datavant.com/.

[27]. Radical RS. RADx-rad Discoveries and Data Coordinating Center (DCC) [Internet]. RADxSM Radical. [cited 2022 Mar 8]. Available from: https://www.radxrad.org/.

[28]. Data privacy principles: Full breakdown of the 7 principles [Internet]. [cited 2022 Mar 11]. Available from: https://www.invisibly.com/learn-blog/data-privacy-principles.

[29]. Art. 5 GDPR - Principles relating to processing of personal data [Internet]. GDPR. eu. 2018 [cited 2022 Mar 11]. Available from: https://gdpr.eu/article-5-how-to-process-personal-data/.

[30]. Warren SD, Brandeis LD, The Right to Privacy [Internet], Harvard Law Review. (1890) 193, 10.2307/1321160.

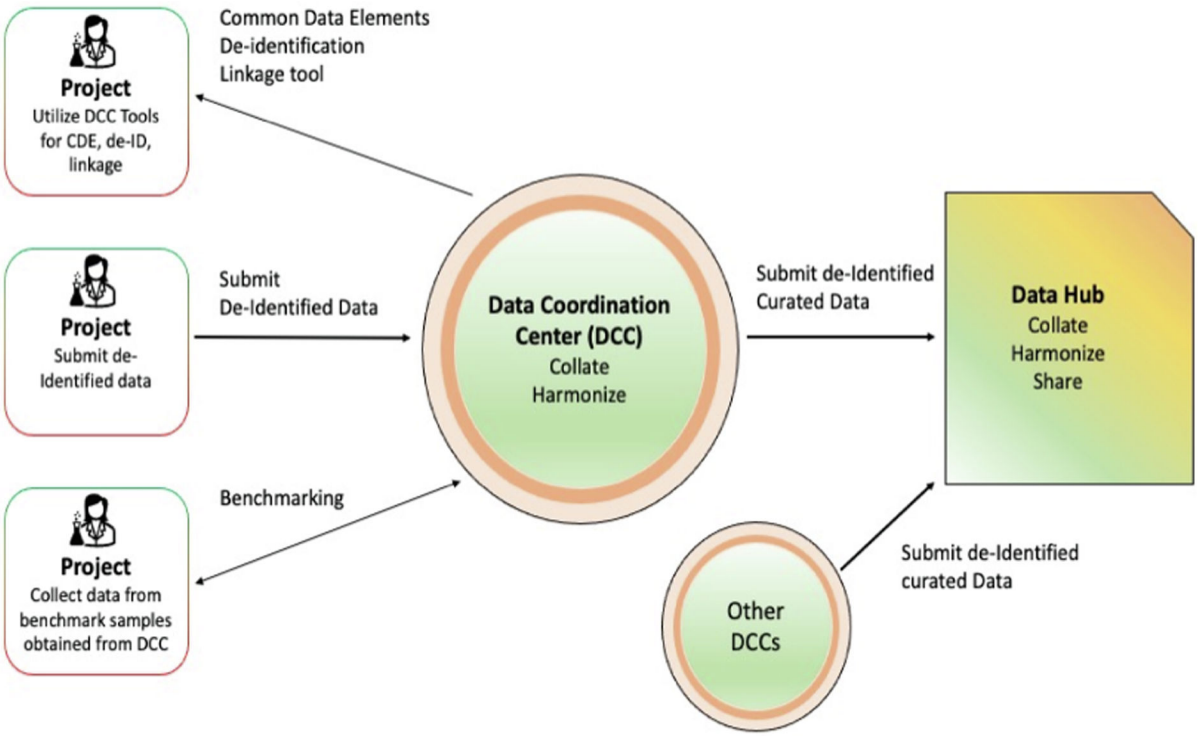

Fig. 1.

A schematic of the central role played by the Data Coordination Center (DCC) in interfacing with the Study investigator's team (e.g., grant awardee) and supporting it to generate consistent, comparable data, by receiving de-identified data from the researchers, collating, harmonizing them across the various projects, and submitting them to the Data Hub.

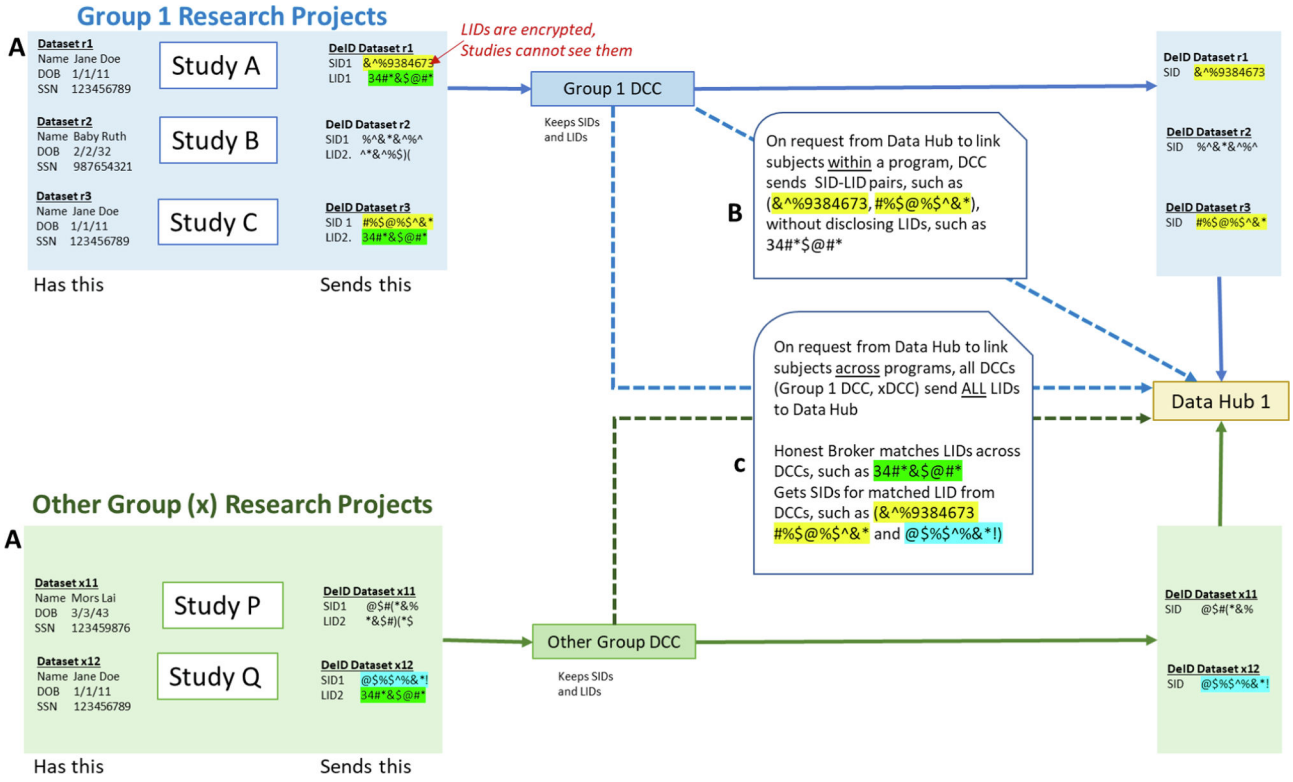

Fig. 2.

An illustration of the data linkage process with Study IDs (SIDs) and Linkable IDs (LIDs). We distribute our executable program for ID generation to get back a pair of (SID, LID) for the Study. The continuous lines above show routine transmission of data from the awardees to the Data Hub via DCC, and the dashed lines show the transmission of linkages when requested by the Data Hub. **(A)** Studies "de-identify" (i.e., remove explicit identifiers) and harmonize the data using Data Coordinating Center (DCC) tools. SID-LID pairs are created for each participant. LIDs are visible only to the DCC. SIDs are unique for each participant and each study, and are not linkable without LIDs. They are visible to the Studies, the DCC and the Data Hub, and could be used by the Study to re-contact the participant after linkage. Both IDs and other data are sent from Studies to the DCC through encrypted communication protocols. The DCC checks "de-identification" and harmonization of data and passes the SID and the other data to the Data Hub, retaining the SID-LID pairs. It may delete other data once the Data Hub receives and approves the data. **(B)** When the Data Hub requests linkage of participants within a set of Studies (e.g., Group 1), the Group 1 DCC will utilize LIDs (highlighted in green) across the studies it coordinates, and will send the SID-SID pairs/sets (highlighted in yellow) of linked participants to the Data Hub. **(C)** When the Data Hub requests linkage of participants across sets of Studies (e.g., Group1 Research Project and another Group(x)'s Project), Group 1's DCC and the other Group(x)'s DCCs send all LIDs to the Data Hub, which acts as the honest broker to determine a list of pairs of matched participants across programs. The Data Hub submits a list of matched LIDs (highlighted in green) to both DCCs and requests that they return the SIDs (highlighted in yellow and blue) that correspond to these LIDs. Note: The seeds for SIDs are kept at the DCCs, whereas the

seed for LIDs is kept at the Data Hub. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

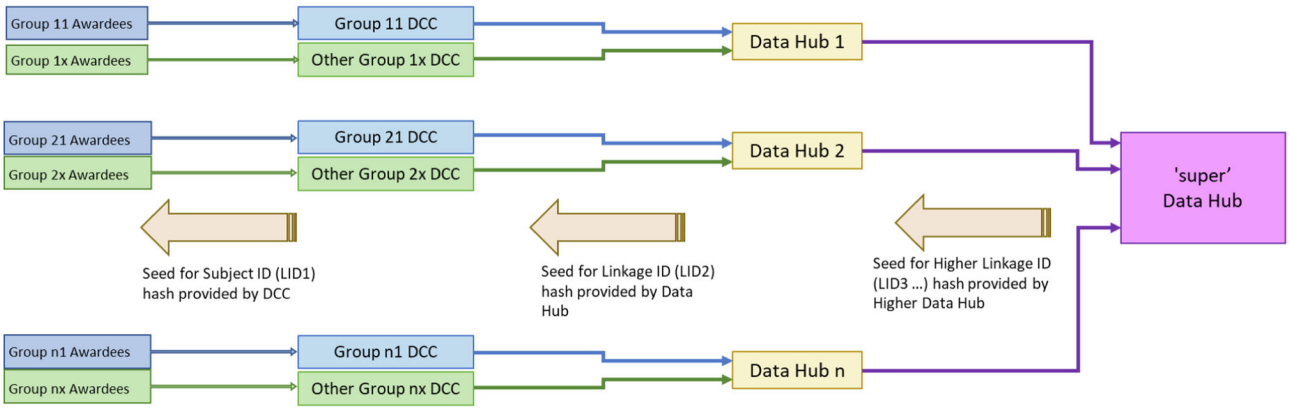

Fig. 3.

Scaling of the single data hub shown in Fig. 1 to privacy protected record linkages across data hubs can be readily extended using the same principles. Fig. 1 details the mechanics associated with an equivalent to Awardees and DCCs reporting into a Data Hub 1 above. It should be noted that the seed for hashing the Subject IDs (LID1s) is provided and maintained by the DCC, while the seed for linkages (LID2s) is provided by the Data Hub. Similarly, for multiple Data Hubs reporting to a ‘super’ Data Hub, the seed for such linkage (LID3s, etc.) should be provided by the ‘super’ Data Hub (*highest level).

Table 1

Information known to each party: a Participant, a Study, the DCC, and the Data Hub.

Information	Participant	Study	DCC	Data Hub
Personal Identifier Information (e.g., name, date of birth, and/or social security number)	Y	Y	N	N
Other Study Data (e.g., measurements, lab values)	N	Y	Y (in “de-identified” form, and can be deleted after successful submission and data approval by the Data Hub)	Y (in “de-identified” form)
Study ID - SID (one per study)	Y (as needed, e.g., to use in a study portal)	Y (for the specific study only)	Y	Y
Linkable ID - LID	N	N	Y (only the DCC has this)	N (unless it requests linkage of data across Studies coordinated at different DCCs, when it will receive lists of LIDs to match from the DCCs)
Linkage results	N	N	N (but gets to know which participants have records in other Studies when linkage happens)	Y (gets pairs of SIDs for the same patient for linked Studies; and learns the LID for participants matched across DCCs)