Budin-Ljøsne, I., Burton, P., Isaeva, J., Gaye, A., Turner, A., Murtagh, M. J., ... Harris, J. R. (2015). DataSHIELD: An Ethically Robust Solution to Multiple-Site Individual-Level Data Analysis. Public health genomics, 18(2), 87-96. 10.1159/000368959

DataSHIELD: an ethically robust solution to multiple site individual-level data analysis

Isabelle Budin-Ljøsne[1], Paul Burton [2], Julia Isaeva[1], Amadou Gaye[2], Andrew Turner[2], Madeleine J Murtagh[2], Susan Wallace[3], Vincent Ferretti[4], and Jennifer R. Harris[1]

[1] Norwegian Institute of Public Health, Division of Epidemiology, Department of Genes and Environment, P.O. Box 4404 Nydalen, NO-0403 Oslo, Norway

[2] University of Bristol, School of Social and Community Medicine, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom

University of Leicester, Department of Health Sciences, Leicester, LE1 7RH, United Kingdom

[3] University of Leicester, Department of Health Sciences, Adrian Building, University Road, Leicester UK, LE1 7RH

[4] Ontario Institute for Cancer Research MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario, Canada, M5G 0A3

Corresponding author:

Isabelle Budin-Ljøsne

Email: isabelle.budin.ljosne@fhi.no

Phone number: +47 21 07 83 02

Fax number: +47 21 07 82 52

KEYWORDS: Data sharing, IRB review, ethics, DataSHIELD, biobank, epidemiological research, statistical analysis

ABSTRACT

Background

DataSHIELD has been proposed to facilitate the co-analysis of individual-level data from multiple studies without physically sharing the data. In a previous paper, we investigated whether DataSHIELD could protect participant confidentiality in accordance with UK law. In this follow-up paper, we investigate whether DataSHIELD can address a broader range of ethics-related data sharing concerns.

Methods

Ethics-related data sharing concerns of IRBs, ethics experts, international research consortia and research participants were identified through a literature search and systematically examined at a multidisciplinary workshop to determine whether DataSHIELD proposes mechanisms which can address these concerns.

Results

DataSHIELD addresses several ethics-related data sharing concerns related to privacy, confidentiality, and the protection of the research participant's rights while sharing data and after the data have been shared. The data remain entirely under the direct management of the study that collected them. Data processing commands are strictly supervised and the data are queried in a protected environment. Issues related to the return of individual research results when data are shared are eliminated as the responsibility for return remains at the study of origin.

Conclusion

DataSHIELD can provide an innovative and robust solution for addressing commonly encountered ethics-related data sharing concerns.

INTRODUCTION

Vast amounts of data are needed to study the causes of disease and elucidate interactions between genes and environment [1]. Building enriched datasets typically involves integrating data from diverse sources, including clinical care, health registries and research data, and often includes transnational data sharing [2]. Such data sharing is increasingly demanded by research funders as a way to accelerate scientific discovery and maximize the economic returns on research data [3-5]. Much of the data sharing that has taken place in international consortia studying genetics and disease has occurred at the aggregate or summary-level for the conduct of meta-analyses [6]. Sharing summary-level data offers more data security than sharing individual-level data but does not offer the analytical flexibility and precision that can be achieved when sharing individual-level data. For instance, summary statistics often fail to convey all of the information held in the individual-level raw data or may not suffice to extend exploration of significant findings. In comparison, sharing individual-level data from local study sites offers much greater analytical flexibility, and sometimes increased precision, because the individual-level data can be pooled and analysed directly. However, it is ethically more challenging because individual-level data may contain sensitive information about the individual's health, lifestyle, genotype or socio-demographic factors that potentially can be used to identify these individuals or provide extensive insight into their private life. Accordingly, mechanisms are typically put in place when sharing data to safeguard against re-identification, prevent potential data misuses and protect privacy and confidentiality. Such mechanisms include both technical (e.g. data coding, password protected access, use of off-site broker with key, limitations on publishable sample size) and administrative (e.g. data access agreements, confidentiality clauses) solutions [7]. However, they often place severe limitations on data sharing, can require considerable administrative effort and do not always

sufficiently address concerns surrounding data sharing. For instance, even if data access agreements are established for a data sharing collaboration, it can prove difficult to control what happens with the data once they are transferred to another site [8-9].

It is will all these considerations in mind that our international team of researchers proposed DataSHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual levEL Databases) [10]. The objective of DataSHIELD is to facilitate the co-analysis of data with all the benefits of individual-level analysis while recognising and finding alternatives that address the major ethical concerns that usually accompany individual-level data sharing. DataSHIELD is being developed by the Data to Knowledge (D2K) Research Group at the University of Bristol under the umbrella of the FP7 collaborative project BioSHaRE (Biobank Standardisation and Harmonisation for Research Excellence in the European Union) [11].

In a previous paper, we investigated whether DataSHIELD could appropriately protect participant confidentiality according to UK legal standards [12]. The paper concludes that DataSHIELD reaches UK standards of protection for the sharing of biomedical data, and calls for the investigation on whether DataSHIELD can satisfy other legal and ethics review requirements, also outside of the UK. In this follow-up paper, we investigate whether DataSHIELD can address a broader range of ethics-related data sharing concerns of Institutional Review Boards (IRBs), ethics experts, international research consortia and research participants, independently on whether there are encountered in the UK or in other countries. It should be noted that this second paper primarily focuses on ethics-related data sharing concerns and does not encompass the analysis of legal requirements.

What is DataSHIELD?

DataSHIELD is an analytical tool that enables the co-analysis of individual-level data from multiple studies or sources without physically transferring or sharing the data and without providing any direct access to individual-level data [13-15]. DataSHIELD can be used to run the same kind of analyses as with any other statistical tool. For instance, DataSHIELD can be used to produce a table showing the age distribution of patients in several studies in percentages; or to analyse variables providing information about age (x1) and smoking habit (x2) with the objective to predict a risk of cancer outcome (y). The range of possible analyses in DataSHIELD is outlined in the DataSHIELD wiki (16).

Figure 1 illustrates how the traditional analytical workflow is reversed under DataSHIELD. Rather than bringing the data to the analyses, the analyses are brought to the data. Individual-level data are never transferred away from the local study computers; parallel data analyses commands are instead simultaneously brought to bear on the individual-level data at each local site involved in the collaboration. Through iterative computational processing, the only information that is transferred back and forth between the local sites holding their data and the analysis centre are the analytical commands and the resultant non-identifying statistical estimates and summary parameters generated from those commands.

As described in Figure 2, DataSHIELD is primarily used for co-analysis of data when each data source contains the same variables (e.g. age, sex, blood pressure) on different individuals (this is called horizontal partitioning) [17]. DataSHIELD is also being developed for co-analysis of data when different data sources (e.g. a cohort study, a hospital record, a registry) report different variables on the same individuals (this is called vertical partitioning). This paper focuses solely on horizontal partitioning which has recently been implemented as an open-source software application and is therefore likely to be encountered by ethics committees, IRBs and other governance boards.

What is needed to use DataSHIELD?

The use of DataSHIELD requires the establishment of a specific IT-environment which includes a central analysis computer, OPAL database servers [18], the open source software for statistical computing R [19] and the DataSHIELD R packages [10]. Both Opal software and DataSHIELD R packages are open source and freely available to the research community. The Opal servers are installed inside the firewall at the local study sites of all the collaborating studies. Other requirements for co-analysis of data under DataSHIELD do not differ from conventional approaches with respect to preparatory activities and include checking that governance stipulations allow the data to be used for the specified project, identifying the variables to use from the different studies, harmonising the measures to be analysed and de-identifying the data to be shared from each of the local datasets.

METHOD

In August 2012, we organized a multidisciplinary workshop gathering biostatisticians, epidemiologists, sociologists, lawyers and ethicists, all involved in the development of DataSHIELD and members of the BioSHaRE project [11]. Before the workshop, a literature search was conducted in Pubmed, Google Scholar and Internet using the combination of the search terms [data sharing] and [ethics] and/or [concerns] and/or [experiences] to identify common ethics-related data sharing concerns of IRBs, ethics experts, international research consortia and research participants. Based on the results from the literature search, a list of commonly encountered ethics-related data sharing concerns was set up and distributed at the workshop. Then, the workshop members conducted a systematic examination of these concerns by identifying and describing the mechanisms in DataSHIELD that may or may not address each of the concerns listed. The objective was to determine whether each concern: (i) could be solved or ameliorated by DataSHIELD; (ii) could be created or made worse by

DataSHIELD; (iii) were independent of DataSHIELD, and so could not be ameliorated by DataSHIELD, but equally was no more of a problem for DataSHIELD than for any other form of data sharing or co-analysis. The discussions at the workshop also encompassed a range of technical statistics/IT considerations, and legal, professional, and societal issues (e.g. related to the appropriate identification of intellectual property and contribution), but this paper focuses solely on key issues from the perspective of ethical and governance boards.

RESULTS

1. Main ethics-related data sharing concerns

Our literature search revealed that ethics-related data sharing concerns are primarily related to 1) the protection of the privacy and confidentiality of the data, 2) the protection of the research participants' rights when data are shared, and 3) what may happen to the data after they have been shared. These concerns are described below and summarized in Table 1.

Concerns related to the protection of the privacy and confidentiality of the data

A major concern of IRB members [20-22], ethics experts [9, 23-28], members of international research consortia [29-31] and research participants [32-36] is that the privacy and confidentiality of the data may be breached when the data are shared, potentially leading to making the participants' specific health risks public. For instance, datasets may accidentally disclose sensitive information, even when they have been modified to include only non-identifiable information, because external investigators are able to link the information in the dataset with information in other publicly available datasets to re-identify individuals [37-39] or because summary data may unexpectedly be found to convey more information that had previously been believed [40-41]. Similarly, a researcher may deliberately violate the terms of

the informed consent and share sensitive data that should not be shared with other investigators outside of the study of origin [42]. The security of the data can also be jeopardized if the individual-level datasets are hacked or copied when physically transferred to a central computing unit for analysis [43].

Concerns related to the protection of the research participants' rights

Several concerns arise in data sharing collaborations regarding the protection of the research participants' rights. First, it is often difficult for researchers to know whether data sharing is compatible with the terms of the original consent [29-31, 44]. This is primarily because many consent forms, particularly those collected some decades ago do not explicitly mention data sharing at all [45]. Second, it is often difficult for researchers to ensure that the research participants' right to withdraw from a study at "any time and without any conditions" (as usually formulated in consents) and the right to require that personal data be deleted and removed from the research databases are sufficiently protected when the data are shared multiple times across studies and managed by others [44]. To address this issue recent versions of informed consents are often modified to explain that data cannot be withdrawn and deleted once they have been physically distributed for analysis [46]. This approach may seem to solve the issue of withdrawal but in practice it restricts the individual's right to withdraw as this right then only applies if the data are not shared. Third, it is often difficult for researchers to know how to handle the feedback of individual research results produced through data sharing to research participants. Although the issue of whether individual research results, in particular from genetic and genomic research, should be returned to research participants is still much debated, several contemporary opinions and guidelines favour return of certain results under specific circumstances [47-51]. Providing such results may not be problematic when the data are processed at the site of the study of origin but this

can become much more complicated when the data are shared. Namely, which investigator is responsible for returning individual research results to participants: the researcher of the original study or the researcher who actually generated the relevant results having gained access to the data at a later time point [51]?

Post-data sharing concerns

Protecting the data and the research participants' rights after the data have been shared is another key concern. For instance, who is responsible for ensuring that data are appropriately stored and curated into the future and who ensures that they are accessed only by those who have proper authorization, if secondary access is awarded to a research group that is then wound up, for example because its leader retires [52]. Although codes of conduct have recently been proposed to help pave the way for a common set of data sharing principles [53,54] and recommendations have been forwarded for the establishment of international governance models when sharing data [55], there is currently no standard protocol to help guide the allocation of complementary governance responsibilities to different research groups (for example, the original data generators and secondary users) or to indicate precisely what these responsibilities may entail [44,56].

Properly addressing the ethics-related data sharing concerns described above is often burdensome and difficult for researchers who have certainly not sought these formal responsibilities. For instance, the more the data are shared, the more difficult it becomes for the investigator of the original study or the biobank which collected the data to monitor and control how the data are handled by others and to properly assess potential risks related to the sharing of those data. This is primarily because the level of risk is a function of the full data environment -- the datasets and the available technologies -- and not just of the dataset alone [9]. Furthermore, having full control regarding the fate of the data over time requires

resources that are often non-existent or scarce [52]. For instance, research collaborations are normally set up for a limited period of time. What happens to the data after the collaboration has ended and how they are to be protected from potential misuses is rarely made explicit and is often unclear [52].

2. How does the DataSHIELD approach address ethics-related data sharing concerns?

DataSHIELD has a number of characteristics that provide solutions to several of the ethics-related data sharing concerns described above. Primarily four sets of mechanisms apply in DataSHIELD to protect the privacy and confidentiality of the data. First, the individual level data are never physically shared or transferred but are instead queried locally. This has positive implications for many of the concerns normally encountered when sharing data as summarized in Table 1. For instance, concerns regarding the protection of the research participants' right to withdraw data from shared datasets become non-existent as the data never leave the local study sites and can easily be removed or destroyed locally. This also allows the local sites to ensure that the research use complies with existing consents. Similarly, returning individual research results to research participants is a non-issue under DataSHIELD because co-analysis in DataSHIELD never produces explicit individual-level research results. This is because, although the contribution of the data from each individual is properly included in every analysis, that contribution is always merged with the equivalent contributions of all of the other participants of that same study before the information driving the overall analysis is transmitted from the study to the analysis centre. This means that individual results are invisible to the statistician coordinating the central analysis and cannot even be inferred by anybody outside the original study itself. One may ask whether designing a system that prevents the return of individual research results to participants is acceptable at a time when such return is increasingly recommended by commentators [47-51]. However, the

decision to use DataSHIELD implies that the return of results has been properly discussed prior to analysis and that the research participants endorse the return policy that applies for them.

Second, each DataSHIELD command systematically goes through a 3-level validation process to ensure that it does what it has been designed for and that potential disclosure risks are kept to a minimum. Each command is internally checked and tested by a DataSHIELD developer other than the one who wrote the command, then checked again by an external 'expert' not involved in the development of DataSHIELD, and finally reviewed by the DataSHIELD Advisory Board which discusses whether the command respects the privacy and confidentiality protecting principles of the DataSHIELD platform. The advisory board may request that some changes are made to the command and takes the final decision of approving or rejecting the command. Commands or sequences of commands that are explicitly disclosive are systematically blocked. In addition, special restrictions may be placed on the nature of the output that a particular DataSHIELD command can return. For example, contingency table analyses can only produce tables which contain no cells with counts between 1 and 4, and where necessary these limits can be tailored to reflect specific legislation in the country of origin of the study. Similarly, when graphical representations are used to display the relationship between two variables, heat map plots and contour plots are used rather than standard point-by-point representations. This is because some such points may be disclosive for certain individuals. If disclosure was to occur, the commands that are responsible for the disclosure can be easily identified as all commands that are issued are recorded and it is kept track of who actually issued them. Any accidental disclosure can therefore lead to a suitable warning, and appropriate sanctions can be applied if deliberate maleficence has occurred.

Third, DataSHIELD includes a number of mechanisms to protect the data from any potential external attack. As described earlier, the use of DataSHIELD involves an internet communication between the central analysis computer and the study's Opal servers. Using the internet to exchange data always involves some level of risk and it is impossible to guarantee that no one will, at some point in time, attempt to compromise the security of the data.

To minimize risks, DataSHIELD follows best practice by ensuring that the operating system and software are secure and kept up-to-date to address new and emerging threats [57-58]. In addition, all communication across the internet between the study computers and the analysis centre is encrypted and secured. For instance, web services are accessed through Hypertext Transfer Protocol Secure (HTTPS) and Opal systematically checks the digital signatures of any user [59]. IP address filtering can be configured in the study's firewalls to prevent any other computer than the allowed central analysis one to connect to the Opal servers. Even if someone was to hack in and decrypt the data traffic flowing back and forth between the analysis centre and the local studies, that traffic is deliberately non-disclosive: this being the fundamental basis of DataSHIELD [14].

In some cases, although the main database of a given study may be too sensitive to allow any risk of access via the internet, the subset of data required for a particular analysis under DataSHIELD may not demand such stringent isolation. In such cases, it is possible to place the data to be used in the analysis in a separate database still located behind the firewall of the study. It should however be noted that in cases where the absolute security of the data is of utmost importance then the best practice for data of this kind is for it to be inaccessible from the internet, in which case DataSHIELD is not an appropriate tool to use.

Finally, DataSHIELD is an open source tool. It can be examined and audited by any potential user who can contribute to its future improvement, which means that no one has to take on trust claims that its operations are secure: users can check for themselves.

DISCUSSION

The main ethics-related data sharing concerns relate to the protection of the privacy and confidentiality of the data and the protection of the research participants' rights both while the data are being shared and after the data sharing has taken place. These results are corroborated by findings from and a video ethnography (observation) study of an early DataSHIELD development workshop [15]. In this study, the centrality of concerns about the maintenance of privacy and confidentiality for individual-level data by DataSHIELD developers and would-be users was demonstrated.

Our analysis reveals that many of the most common ethics-related data sharing concerns become non-issues or are greatly alleviated under DataSHIELD. Concerns related to the protection of the research participants' rights are eliminated because the data are never physically shared and therefore remain entirely under the direct management of the study that collected them. Concern related to the protection of the privacy and confidentiality of the data is minimized as the data are never physically accessed by others and key security features are built into DataSHIELD to reduce disclosure risks. This may significantly change the way cross-study analyses are conducted in research collaborations and facilitate the conduct of research projects which otherwise would be difficult to realize due to privacy and security concerns. As an illustration, researchers often need to pool data from diverse sources, for instance medical records, to conduct research investigating the aetiology of disease or mechanisms underlying side effects of medical treatments [60]. Such research is of high value for public health but is often difficult to realize because of the sensitivity of the data held in

medical records. DataSHIELD may provide a useful solution to conduct such research as risks of breaching patient confidentiality by using DataSHIELD, although not entirely eradicated, would be reduced to an "absolute and acceptable minimum" [60].

DataSHIELD may also facilitate the sharing of data that otherwise would not be shared due to intellectual property concerns as it allows for the sharing of information held in the data without having to physically transfer or share the data themselves [60]. Finally, DataSHIELD may facilitate the conduct of research projects which normally are too difficult to realize due to technical constraints. For instance, while data sharing often requires lots of computational capacity when large data files are transferred to a central computer for analysis, such capacity is not needed in DataSHIELD since the data files remain on local study sites and it is only the non-disclosive summary statistics that are passed between studies and the analysis centre, and these are generally very small. The use of DataSHIELD may also improve the quality of co-analysis. Study sites participating in a standard collaboration, for instance conventional meta-analysis, are normally required to run statistical analysis of similar quality and design. This can be difficult to coordinate and police when datasets from numerous sites are used. In DataSHIELD, the same data analysis commands are sent to all local study computers simultaneously. Variations in command quality or design are therefore never encountered.

As explained earlier, DataSHIELD cannot be used in research projects which require producing disclosive summaries (such as point-by-point representations in scatter plots) as such features are blocked in DataSHIELD to protect the confidentiality of the data [60]. However, alternative solutions can be provided, for instance graphical representations without individual data points such as contour plots [60].

A central question is whether analysis in DataSHIELD still qualifies as data sharing per se since the individual-level data are never physically shared but queried at local study sites and only summary statistics are shared. In our previous paper led by Susan Wallace [12], we

suggested that the summary statistics processed in DataSHIELD are anonymous data which could potentially be shared without referral to European data protection principles, thus opening for pan-European use of the data. A similar analysis could indicate whether DataSHIELD can cross internal national borders (i.e. US state or Canadian provincial borders) or international borders. Current practice is that researchers normally do not share individual level data if the consent of the study of origin does not allow for such sharing or does not specifically mention the possibility of data sharing. Such practice is legitimate but limits the possibilities of retrospective research when the consents do not mention data sharing. It can reasonably be argued that the analytical process in DataSHIELD should be considered to be equivalent to meta-level analysis using summary level data (which is normally the standard data sharing practice when the informed consent does not mention or authorise data sharing). However, technological approaches should not be used as a way of circumventing informed consent. Therefore further research is needed to determine whether IRBs and research participants would be comfortable with the use of DataSHIELD in the absence of explicit consent but with the approval of ethics and scientific review bodies.

As an entirely new approach to the joint analysis of data from several studies, DataSHIELD offers some potentially exciting opportunities. We encourage members of IRBs and ethics committees to consider and discuss whether the use of DataSHIELD is consistent with the original intents for use of data as framed in the informed consents of the studies they manage. Similarly, we encourage researchers to consider whether the use of DataSHIELD may be useful in their research collaborations. Feedback from the community on this matter is greatly appreciated.

CONCLUSION

Multiple site individual-level data analysis is increasingly needed to accelerate research discovery but encounters a number of ethical challenges. DataSHIELD offers a new approach to data sharing and is currently being tested in real-life epidemiological projects, including the Healthy Obese Project of the BioSHaRE project [11]. In our previous paper led by Susan Wallace [12], we concluded that DataSHIELD was in compliance with UK standards of protection for the sharing of biomedical data. This new paper demonstrates that DataSHIELD can also address a number of commonly encountered ethics-related data sharing concerns. New commands are being developed in DataSHIELD to address the needs of a variety of collaborations. Further work is needed to investigate whether the use of DataSHIELD is compliant with legal requirements in countries other than the UK.

ACKNOWLEDGMENTS

This work was supported through funds from the European Union's Seventh Framework Programmes ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413; BioSHaRE-EU, grant agreement HEALTH-F4-2010-261433; and Biobank Norway, funded by the Norwegian Research Council (NFR 197443/F50). The development and application of DataSHIELD is also funded under a strategic award from MRC and Wellcome Trust underpinning the ALSPAC project, the Welsh and Scottish Farr Institutes funded by MRC, and BBMRI-LPC (EU FP7, I3 grant).

Reference list

1. Committee on a Framework for Development a New Taxonomy of Disease; National Research Council. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. The National Academies Press. 2011.

2. Burton PR, Hansell AL, Fortier I, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. International Journal of Epidemiology 2009; 38:263-273.

3. National Institutes of Health. NIH Data Sharing Policy. 2007. http://grants.nih.gov/grants/policy/data_sharing/

4. Organisation for Economic Co-operation and Development. OECD principles and guidelines for access to research data from public funding. 2007. http://www.oecd.org/sti/sci-tech/38500813.pdf

5. Wellcome Trust. Wellcome Trust Data Sharing Policy. 2013. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/

6. Hindorff LA, Sethupathy P, Junkins HA, et al: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences 2009 June 9, 2009;106(23): 9362-7.

7. Reiter JP, Kinney SK: Sharing confidential data for research purposes: a primer. Epidemiology 2011; 22:632-635.

8. Pearce N, Smith AH: Data sharing: not as simple as it seems. Environmental Health 2011; 10:107.

9. Kaye J, Heeney C, Hawkins N, el al: Data sharing in genomics--re-shaping scientific practice. Nature Review Genetics 2009; 10:331-335.

10. Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual levEL Databases (DataSHIELD). URL: http://datashield.org/

11. Biobank Standardisation and Harmonisation for Research Excellence in the European Union BioSHaRE-EU. 2013. URL: https://www.bioshare.eu/

12. Wallace SE, Gaye A, Shoush O, Burton PR. Protecting Personal Data in Epidemiological Research: DataSHIELD and UK Law. Public Health Genomics. 2014 Mar 28.

13. Jones E, Sheehan N, Masca N, el al: DataSHIELD – shared individual level analysis without sharing the data: a biostatistical perspective. Norsk Epidemiologi 2012; 21 (2), 231-239.

14. Wolfson M, Wallace SE, Masca N, et al: DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. International Journal of Epidemiology 2010; 39:1372-1382.

15. Murtagh MJ, Demir I, Jenkings KN, et al: Securing the data economy: translating privacy and enacting security in the development of DataSHIELD. Public Health Genomics 2012; 15:243-253.

16. Gaye A, Wilson R, Turner AJ. DataSHIELD wiki - Packages and Functions available. URL: https://wikis.bris.ac.uk/display/DSDEV/List+of+Packages+and+functions+currently+available

17. Doiron D, Burton P, Marcon Y, et al: Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerging Themes in Epidemiology 2013; 10(1):12

18. OPAL. 2013. http://obiba.org/node/63

19. R software. 2013. http://www.r-project.org/

20. Lemke AA, Smith ME, Wolf WA, el al: Broad data sharing in genetic research: views of institutional review board professionals. IRB. 2011; 33(3):1-5.

21. Wolf LE, Catania JA, Dolcini MM, et al: IRB Chairs' Perspectives on Genomics Research Involving Stored Biological Materials: Ethical Concerns and Proposed Solutions. Journal of Empirical Research on Human Research Ethics. 2008 ; 3(4):99-111.

22. Kozanczyn C, Collins K, Fernandez CV: Offering results to research subjects: U.S. Institutional Review Board policy. Accountability in Research: Policies and Quality Assurance 2007; 14(4):255-67.

23. Heeney C, Hawkins N, de VJ, el al: Assessing the privacy risks of data sharing in genomics. Public Health Genomics 2011; 14:17-25.

24. Knoppers BM, Dove ES, Litton JE, et al: Questioning the limits of genomic privacy. American Journal of Human Genetics 2012; 91:577-578.

25. Global alliance to create standards for sharing genomic data: group supports simplifying system for searches, but privacy a concern. Am J Med Genet A. 2013 Sep;161(9):xi. doi: 10.1002/ajmg.a.36168.

26. McEwen JE, Boyer JT, Sun KY. Evolving approaches to the ethical management of genomic data. Trends Genet. 2013 Jun;29(6):375-82.

27. Brenner SE. Be prepared for the big genome leak. Nature. 2013 Jun 13;498(7453):139.

28. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Investig Med. 2010 Jan;58(1):11-8.

29. McGuire AL, Basford M, Dressler LG, et al: Ethical and practical challenges of sharing data from genome-wide association studies: the eMERGE Consortium experience. Genome Research 2011; 21:1001-1007.

30. Peppercorn J, Shapira I, Deshields T, et al: Ethical aspects of participation in the Database of Genotypes and Phenotypes of the National Center for Biotechnology Information: The Cancer and Leukemia Group B Experience. Cancer 2012; 118:5060-5068.

31. Zink A, Silman AJ: Ethical and legal constraints on data sharing between countries in multinational epidemiological studies in Europe report from a joint workshop of the European League Against Rheumatism standing committee on epidemiology with the "AutoCure" project. Annals of the Rheumatic Diseases 2008; 67:1041-1043.

32. Lemke AA, Wolf WA, Hebert-Beirne J, et al: Public and biobank participant attitudes toward genetic research participation and data sharing. Public Health Genomics. 2010; 13(6):368-77.

33. McGuire AL, Hamilton JA, Lunstroth R, et al: DNA data sharing: research participants' perspectives. Genetics in Medicine 2008; 10(1):46-53.

34. Oliver JM, Slashinski MJ, Wang T, et al: Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. Public Health Genomics 2012; 15(2):106-14.

35. Trinidad SB, Fullerton SM, Bares JM, et al: Genomic research and wide data sharing: views of prospective participants. Genetics in Medicine 2010; 12(8):486-95.

36. Burstein MD, Robinson JO, Hilsenbeck SG, et al. Pediatric data sharing in genomic research: attitudes and preferences of parents. Pediatrics. 2014 Apr;133(4):690-7.

37. Gymrek M, McGuire AL, Golan D, et al: Identifying personal genomes by surname inference. Science 2013; 339:321-324.

38. McGuire AL, Gibbs RA. Genetics. No longer de-identified. Science. 2006 Apr 21;312(5772):370-1.

39. El Emam K, Buckeridge D, Tamblyn R, et al. The re-identification risk of Canadians from longitudinal demographics. BMC Med Inform Decis Mak. 2011 Jun 22;11:46.

40. Homer N, Szelinger S, Redman M, et al: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics 2008; 4:e1000167.

41. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science. 2004 Jul 9;305(5681):183.

42. Mello MM, Wolf LE. The Havasupai Indian tribe case--lessons for research involving stored biologic samples. N Engl J Med. 2010 Jul 15;363(3):204-7.

43. Erlich Y, Narayanan A2. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014 Jun;15(6):409-21.

44. Kaye J, Hawkins N. Data sharing policy design for consortia: challenges for sustainability. Genome Med. 2014 Jan 29;6(1):4.

45. Tassé AM, Budin-Ljøsne I, Knoppers BM, et al: Retrospective access to data: the ENGAGE consent experience. European Journal of Human Genetics 2010; 18(7):741-5.

46. Organisation for Economic Co-operation and Development (OECD). OECD Guidelines on Human Biobanks and Genetic Research Databases. 2009. URL: http://www.oecd.org/science/biotech/44054609.pdf

47. Fabsitz RR, McGuire A, Sharp RR, et al: Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. Circulation: Cardiovascular Genetics 2010; 3:574-580.

48. Knoppers BM, Deschenes M, Zawati MH, et al: Population studies: return of research results and incidental findings Policy Statement. European Journal of Human Genetics 2013; 21:245-247.

49. Wolf SM, Crock BN, Van NB, et al: Managing incidental findings and research results in genomic research involving biobanks and archived data sets. Genetics in Medicine 2012; 14:361-384.

50. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013 Jul;15(7):565-74.

51. Knoppers BM, Joly Y, Simard J, Durocher F. The emergence of an ethical duty to disclose genetic research results: international perspectives. Eur J Hum Genet. 2006 Nov;14(11):1170-8.

52. Budin-Ljosne I, Isaeva J, Maria KB, et al: Data sharing in large research consortia: experiences and recommendations from ENGAGE. European Journal of Human Genetics 2013.

53. Knoppers BM, Harris JR, Tasse AM, et al: Towards a data sharing Code of Conduct for international genomic research. Genome Medicine 2011; 3:46.

54. Knoppers BM, Harris JR, Budin-Ljøsne I, Dove ES. A human rights approach to an international code of conduct for genomic and clinical data sharing. Hum Genet. 2014 Jul;133(7):895-903.

55. Caulfield T, McGuire AL, Cho M, et al: Research ethics recommendations for whole-genome research: consensus statement. PLoS Biology 2008; 6:e73.

56. Boyd D, Crawford K. Critical Questions for big data. Information, Communication & Society. 2012; 15:5, 662-679

57. International Epidemiology Association Guidelines for Proper Conduct in Epidemiologic Research. 2007. http://ieaweb.org/good-epidemiological-practice-gep/.

58. Information Commissioner. Anonymisation: Managing data protection risk code of practice; in Commissioner I (ed). Wilmslow, Cheshire, Information Commissioner's Office. 2012.

59. Opal Configuration Guide. 2014. http://wiki.obiba.org/display/CAG/Home

60. Gaye A, Marcon Y, Isaeva J, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. International Journal of Epidemiology 2014;In press.

Table 1 – Budin-Ljøsne et al: Common ethics-related data sharing concerns and how they are addressed by DataSHIELD

Data sharing concerns	How they are usually addressed	How they are addressed in DataSHIELD
Protection of the privacy and confidentiality of the data		
Breaches of privacy and confidentiality of the data	• Technical mechanisms (e.g. data coding, password protected access, use of off-site broker with key, limitations on publishable sample size) • Administrative mechanisms (e.g. data access agreements, confidentiality clauses)	In addition to standard technical and administrative mechanisms: • Individual-level data never physically shared with researchers outside of the study of origin • 3-level testing of commands for risks of disclosure • Output restrictions to impede return of possibly identifiable results • New subject's identifiers automatically generated by Opal. Original subject's identifiers assigned by studies never exposed and stored securely in a distinct database in Opal
Risk of residual or inferential disclosure	• Standard statistical disclosure methodologies	• Standard statistical disclosure methodologies • Any disclosure can be easily identified, investigated and managed
Risk of hacking in via a portal to the internet	• No standard solution. If the absolute security of a given data set is of utmost importance then best practice is for it to be inaccessible from the internet.	• Moving the data for the DataSHIELD analysis to a separate database behind the study's firewall and using DataSHIELD via an Opal server
Protection of the research participants' rights		
Data sharing according to the terms of the original consent	• Necessary ethico-legal and data access approvals required	• Necessary ethico-legal and data access approvals required
Complexity of guaranteeing the right to withdraw data from shared datasets	• Clause in informed consent that the data cannot be withdrawn once they are shared	• Individual-level data never shared, can therefore be withdrawn/deleted locally
Complexity of returning individual results to research participants	• Variety of policies: from no return of results to some return of validated clinically useful results	• Individual research results are never produced, so no results to return. The exploration, identification and return of potentially relevant individual-level results remain sole responsibility of the local study that originally collected the data.
Post-data sharing concerns		
Complexity of protecting the data and the research participants' rights once the data have been shared	• No standard solution	• Individual-level data are never physically shared. All aspects of the ongoing management of data and research participants' rights in relation to those data remain with the local study.

Figure 1 – Budin-Ljøsne et al.

DataSHIELD analytical flow

Figure 2 – Budin-Ljøsne et al.

Horizontal versus vertical partitioning