Protecting Personal Data in Epidemiological Research: DataSHIELD and UK Law

Susan E. Wallace\textsuperscript{a} Amadou Gaye\textsuperscript{b} Osama Shoush\textsuperscript{a} Paul R. Burton\textsuperscript{b}

\textsuperscript{a}Department of Health Sciences, University of Leicester, Leicester, and \textsuperscript{b}School of Social and Community Medicine, University of Bristol, Bristol, UK

\textbf{Abstract}

\textbf{Background:} Data from individual collections, such as biobanks and cohort studies, are now being shared in order to create combined datasets which can be queried to ask complex scientific questions. But this sharing must be done with due regard for data protection principles. DataSHIELD is a new technology that queries nonaggregated, individual-level data in situ but returns query data in an anonymous format. This raises questions of the ability of DataSHIELD to adequately protect participant confidentiality. \textbf{Methods:} An ethico-legal analysis was conducted that examined each step of the DataSHIELD process from the perspective of UK case law, regulations, and guidance. \textbf{Results:} DataSHIELD reaches agreed UK standards of protection for the sharing of biomedical data. All direct processing of personal data is conducted within the protected environment of the contributing study; participating studies have scientific, ethics, and data access approvals in place prior to the analysis; studies are clear that their consents conform with this use of data, and participants are informed that anonymisation for further disclosure will take place. \textbf{Conclusion:} DataSHIELD can provide a flexible means of interrogating data while protecting the participants’ confidentiality in accordance with applicable legislation and guidance.

\textbf{Key Words}

Biobank · Data sharing · Epidemiological research · Ethics · Law · Policy · Statistical analysis

\textbf{Introduction}

Large collections of biomedical data are becoming increasingly available for use by the research community. Data resources from biobanks, cohort studies, and healthcare systems are being combined and shared in order to ‘… accelerate health research and translation.’ [1]. These collaborations present administrative, ethical, and legal challenges as the confidentiality of the identifiable data collected from participants must be protected and its sharing controlled through applicable legislation, guidance, ethics review, and data access mechanisms [2]. While such controls rightly provide protections, there is a perception that they may also limit access to data and may hinder the research process [3]. Yet the availability of large datasets and the moves to combine data hold the promise of being able to ask and answer increasingly complex questions that will significantly benefit human
health, questions that cannot be answered by one study alone [4]. In addition, some funding agencies require researchers to reuse and exploit the data coming from funded studies, for example, in the area of genomic research data and biobanking [5]. This tension between protecting data confidentiality and enabling data use throws up varying challenges [6], some of which are addressed by new technological initiatives for the sharing of data across research studies.

When data from several studies, or other sources, are to be coanalysed, 2 fundamental approaches exist [1]. Individual-level meta-analysis (ILMA) requires each participating study to contribute individual-level data to a central pooled data repository. These centralised data may then be queried as flexibly and easily as if they had all been collected by one study. As an example, each biobank involved in the German-Austrian Central Research Infrastructure for Molecular Pathology (CRIP) codes specific data, which are then imported into the CRIP In-house Research Database and aggregated [7]. Terms reflecting between-study heterogeneity must typically be incorporated. Individual-level data are variously referred to as ‘individual patient data’ by meta-analysts and as ‘micro-data’ by bioinformaticians [2]. As an alternative, combined analysis may be based on study-level meta-analysis (SLMA). Here, the precise analysis to be undertaken is specified ahead of time. Each study then analyses its own data under the prescribed analysis plan and returns the result of each component of the plan to an analysis centre (AC) which then meta-analyses those results to provide answers across all studies combined.

From an analytic perspective, ILMA is generally to be preferred. Because the analysing statistician has direct access to the data, analysis is completely flexible, and complex data can be explored without a need for a tightly prespecified analysis protocol. Furthermore, some types of analysis can only be undertaken using individual-level data, and ILMA can sometimes be more efficient. On the other hand, an ILMA can raise serious ethico-legal and governance issues. Obviously, if the individual-level data to be shared include unique personal identifiers, such as an NHS number or a social security ID, individuals will be immediately identifiable. ‘Pseudo-anonymised’ or ‘coded’ [8] data protect against identifiability by stripping off identifiers and replacing them with codes that can only be mapped to an individual using a key that is kept confidential by the releasing study. Nevertheless, as more variables are released, there is an increasing risk that in combination they provide enough information to identify some individuals [9–11]. Consequently, many studies have some restriction on the release of individual-level data, and this can hinder ILMA. For example, if the original consent forms, which participants signed in joining a study, implied that their individual-level data would never be shared with researchers external to the study itself, this would preclude any ILMA-based meta-analysis involving analysis elsewhere. Alternatively, an information leaflet implying that individual-level data would never be sent overseas would restrict international ILMA [12].

To circumvent this problem, many research groups use SLMA rather than ILMA. SLMA is well illustrated by a typical genome-wide association study (GWAS). Here, each study undertakes a GWAS only on its own data. The association statistics for each variable (here a SNP) in each study are then shared and submitted to a standard meta-analysis at a designated AC. Crucially, the association statistics are calculated at the level of the study and so can freely be shared amongst collaborative partners, even if there is a formal ethico-legal embargo on sharing individual-level data. Providing a real-world illustration of this approach, the US Food and Drug Administration ‘Sentinel’ Initiative is creating an electronic network that will operate across different data sources in order to monitor medical products approved by the US Food and Drug Administration [13]. Currently, each site transforms subsets of their data into an agreed format, and a coordinating centre then asks each site to run queries and return only summary or aggregated data. No individual-level data leave any participating site, yet information can usefully be shared. Unfortunately, there is a serious scientific drawback to an SLMA. That is, all queries must be prespecifiable so each site can perform the same analyses to produce the same set of summary statistics. This requirement seriously restricts the ability to conduct richer, more exploratory analyses – unless one is willing to repeatedly send out new analytic requests and to wait for the required summary statistics to be returned.

DataSHIELD [14] reflects a radically different approach to both ILMA and SLMA. The data in each repository are analysed at an individual level, but separate and parallelised analyses are undertaken simultaneously in each study. Nevertheless, a typical DataSHIELD analysis is mathematically equivalent to pooling all of the data centrally and carrying out a single analysis [15, 16]. Summary statistics – commonly score vectors and information matrices – are returned by each study to the analyst (fig. 1) at each iteration. But these are nondisclosive; they carry no information about individual participants. And so, under DataSHIELD, although a full individual-level
DataSHIELD and UK Law

Public Health Genomics 2014;17:149–157
DOI: 10.1159/000360255

analysis is enabled, no identifying or sensitive information is physically moved, or even rendered temporarily visible, outside the original study in which the data were collected [17]. In this way DataSHIELD seeks ‘… to elide privacy concerns about sharing data’ while enabling a flexible interrogation of the participating datasets [2]. It is currently being piloted in several projects, including the Healthy Obese Project (HOP), under the Seventh Framework Programme for Research-funded BioSHaRE study [18].

At different times, during a typical DataSHIELD analysis, study data are interrogated in a variety of different ways in order to generate the information that drives the analysis. Different classes of interaction with the data may have differing rules that govern them. In this paper, we present an ethico-legal analysis that explores each step of the DataSHIELD process from the perspective of the ethico-legal framework that applies in the UK. The paper asks what classes of data are processed at each step, and determines whether there is any step at which the data variables or summary statistics that are available outside the original study in which they arose may be used to infer the identity of an individual or anything about his/her characteristics. Critically, we explore whether there is any sense in which personal data, as defined by UK legislation and case law, is potentially revealed. We further ask whether the ways in which DataSHIELD generates, processes, and transmits anonymous summaries are consistent with best practice. We cite recent UK case law and scholarly analyses and argue that DataSHIELD reaches agreed UK standards of protection for the sharing of biomedical data. We discuss limitations to this research and suggest future studies that would allow researchers to improve and enhance DataSHIELD processes.

The DataSHIELD Process

Preliminary Set-Up

As with all joint analyses using SLMA, ILMA, or any other approach, DataSHIELD requires some preliminary steps to prepare the data before analyses can begin (table 1). Any biobank or cohort study site wishing to participate in a DataSHIELD analysis needs to ensure that the appropriate ethics and scientific review and governance approvals that cover that study allow DataSHIELD to be used. If not, additional approval may be needed. Approval will also be needed from the study’s data access committee to obtain the particular set of variables needed for the planned analysis. Then, no joint analysis of any kind is possible unless the data from the different collaborating studies are adequately harmonised [19]. Each study identifies the variables within their own main data repository that are required to generate these harmonised variables. Then, ideally with the support of a central harmonisation unit [20, 21], each study constructs and runs algorithms to generate all of the harmonised variables. Finally, before beginning a DataSHIELD analysis, the data will be appropriately de-identified. Regardless of how sophisticated a secure parallelised analysis may be, there is a high risk of disclosure if the data feeding into the analysis include direct individual identifiers.

At each collaborating centre, a dedicated data computer (DC) is set up with the necessary software and appropriate firewall protections. In principle, this could be the primary data repository for the study in question, but in practice it is much more straightforward to set up, maintain, and secure a new database and server solely for the purpose of the particular set of analyses that are planned. The DC should be set up behind the firewall of the collaborating study and will consist of a server running an OPAL instance, which provides the database environment containing the data to be shared, and an R instance, under which the DataSHIELD analysis can be carried out. OPAL is the core database application for...
biobanks and epidemiological studies that has been developed by OBiBa, an international software development project that builds open source software for biobanks [22]. R is an open source software environment and programming language for statistical computing [23]. The DC is controlled by the collaborating study, and most crucially, the data are only accessible via DataSHIELD. The DC’s information content can be extracted under a valid DataSHIELD analysis, but the data themselves are invisible to anyone outside the study.

Analysis

Once the data are formatted as needed, DataSHIELD allows a researcher at an AC to begin the analysis phase. The analysis can be a single-step command where, for example, the analyst issues a command in the form of a mathematical formula that says ‘produce a table of disease status (diseased or healthy) versus age, for four age groups’. In this single-step case, the DC returns to the AC a summary statistic which consists of a simple $2 \times 4$ contingency table that, with one limitation, can be safely returned by each study to the AC without disclosing sensitive or identifying data; the AC then combines the tables returned by the studies into one table. The one limitation is that any cell containing between 1 and 4 observations may be potentially disclosive, so DataSHIELD prevents such tables from being returned. More complex analyses such as a regression analysis (typically a generalised linear model) which can be used to look at the relationship between an outcome and covariate(s) (e.g. between diabetes status and blood cholesterol levels) can also be run in an iterative process. In such advanced analysis, the AC, which co-ordinates the parallel analyses, sends a command (a model and its first estimates) to the DCs to fit the regression model. The request is then processed within the DC behind the study’s firewall, and a response which consists of summary statistics (for a generalised linear model the summaries are score vectors and information matrices) is transmitted back to the AC. The answers from the studies are then combined by the AC which then sends back updated estimates to the DCs. This iterative process continues until the model’s estimates remain unchanged, producing a result that is identical to one that would have been achieved by harmonising the data at each site, pooling them into one single database and conducting analyses [15, 16].

DataSHIELD Data and Data Protection Principles

There are acknowledged difficulties in sharing samples and data across national borders, due to differing legislative requirements and cultural norms [24]. As discussed, DataSHIELD seeks to bypass these concerns by never sharing individual-level data outside the study in which they were originally collected. DataSHIELD processes the individual-level data but is designed so that the information leaving each study never pertains to the individual participants and is fundamentally nondisclosive. But, in doing all of this, how is DataSHIELD actually pro-
ecting the confidentiality of participant data and respecting the wishes of participants regarding its use?

Step 1 in Table 1 makes clear that governance requirements and participant consents for participating studies must comply or judged to be compatible with this form of scientific research. When piloting DataSHIELD as part of the HOP study, the consent materials from each biobank were reviewed to ensure that the consent language was broad enough to encompass the processing of data within the biobank for research purposes and for the sharing of anonymised data [18]. This review process continues for any new biobank joining the HOP study.

Step 3 in Table 1 ensures that in DataSHIELD no one uses variables that are potentially identifying unless it is absolutely essential to do so. Best practice requires the epidemiologists to protect the confidentiality of research data [25, 26]. For example, it is better to include just the year of birth (or age in years) in a dataset rather than the full date of birth which is more likely to identify an individual. However, if an analysis needs to take account of the age in years and a strong seasonal effect, it may be necessary to use the full date of birth. Although in using DataSHIELD no one will see individual dates of birth, it also follows best practice by not using inherently identifying variables if not necessary.

Even when used in accordance with best practice, there are still a series of specific questions that demand asking and answering in order to indicate whether DataSHIELD meets legally defined criteria for security and confidentiality. First, does DataSHIELD entail the processing of personal data as defined in law and if so, what protections are in place? Second, under DataSHIELD, are individual-level personal data processed comprehensively enough that the resultant summary statistics can be viewed as sufficiently anonymised (nondisclosive) to be viewed as no longer identifiable? Third, does DataSHIELD address concerns regarding a possible breach of a person’s right to privacy? Fourth, can the processing of anonymised data still fall under data protection principles? These questions are addressed in the context of UK legislation and case law.

Is DataSHIELD Processing Personal Data?

In the UK, the Data Protection Act 1998 (DPA 1998) regulates uses of identifiable or ‘personal data’, defined in Section 1(1) as that which relates ‘… to a living individual who can be identified (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of the data controller.’ Eight data protection principles, together with accompanying schedules, detail the ways in which data must be obtained, processed, maintained, and transferred. The principles require that data are:

- used fairly and lawfully,
- used for limited, specifically stated purposes,
- used in a way that is adequate, relevant and not excessive,
- accurate,
- kept for no longer than is absolutely necessary,
- handled according to people’s data protection rights,
- kept safe and secure,
- not transferred outside the UK without adequate protection [27].

Processing ‘sensitive personal data’, which includes information on an individual’s physical or mental health or condition, may require receiving explicit consent for each use of data.

During the preparatory stages of DataSHIELD and an analysis using it, studies will be processing individual-level, and therefore, personal data. During the set-up (Table 1), studies will select and move variables into a separate dataset, and during an analysis, the data will be interrogated to calculate the score vectors and information matrices. But DataSHIELD is specifically designed to ensure that any processing of personal data is conducted within the contributing study behind their firewall and is, therefore, covered by the governance approvals obtained by the study prior to analysis. Studies will need adequate data security protocols [28] as there is always a potential for someone to ‘hack’ into any study database; however, no additional precautions are needed due to using DataSHIELD. The information returned following DataSHIELD requests consists of aggregated data (summary statistics) which are nondisclosive. This means that even if a hacker ‘intercepts’ the response sent back to the user, there will still be no breach of confidentiality.

Are the Data Produced through a DataSHIELD Analysis Anonymised Data?

We argue that all other data produced in a DataSHIELD analysis is anonymised data. According to Recital 26 of the European Data Protection Directive (95/46/EC) [hereafter ‘Directive’], ‘… the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable.’ Therefore, anonymised data can be shared without referral to the data protection principles previously discussed. But what does ‘rendered anonymous’ mean? It is not defined in the Directive or the DPA 1998. Varying definitions exist; we
have chosen the interpretation used by the UK Information Commissioner: ‘…data that does not itself identify any individual and that is unlikely to allow any individual to be identified through its combination with other data.’ [26]. Several recent UK court cases have helped to define and clarify the acceptable use of anonymised, and in particular statistical, data [29, 30].

The case Common Services Agency v Scottish Information Commissioner [31] centred on the Common Services Agency’s (CSA) refusal of a request made under the Freedom of Information (Scotland) Act 2002 for details of incidents of childhood leukaemia in 2 specific regions of Scotland. The CSA argued that the requested statistics could not be released, ‘… because there was a significant risk of the indirect identification of living individuals due to the low numbers resulting from the combination of the rare diagnosis, the specified age group and the small geographic area.’ [31]. Therefore, these statistics were personal data under the DPA 1998 and exempt from disclosure under Section 38 of the Freedom of Information (Scotland) Act 2002. The Scottish Information Commissioner, on the other hand, argued that the data could be ‘Barnardised’ [32] or altered to protect confidentiality and then disclosed.

The judgement centred on 2 questions. First, were the Barnardised data anonymised data that could not identify individuals and were, therefore, not personal data as defined by law? Barnardisation is a recognised statistical disclosure control method where data are modified to disguise information, while not affecting its usefulness in analyses. Analogous methods include suppression, micro-aggregation and removing variables [26] and are used widely in order to share data, such as census data, globally [33]. The second question was whether the Barnardised data remained personal data, because the CSA held the links that could re-establish identity.

The Lords agreed that if it was shown that individuals could not be identified from the Barnardised data, then they were not personal data, data principles did not apply, and they could be disclosed. This was true even if the person controlling the data held the personal data as well. Baroness Hale summed up the argument by stating that,

This ‘common sense’ approach avoided the potentiality of making all anonymised data personal simply because the data controller held the link. The leading judgement in the case, given by Lord Hope, additionally focused on the possibility that someone other than the data controller would have ‘other information’ as well as the Barnardised data, and on the duties of the data controller to judge whether the Barnardisation process had been sufficient to anonymise the data. As a result, the case was referred back to the Scottish Information Commissioner who was asked to review the Barnardisation process to see if it would adequately protect the individuals involved. A subsequent case, Department of Health v Information Commissioner [34], where an agency refused to disclose statistical data about the conduct of abortions because of fears the data would identify individuals, confirmed the findings in CSA.

These judgements have set a legal precedent in the UK and can be seen to clarify the requirements DataSHIELD analysis data must meet in order to be classed as ‘rendered anonymous’ and not subject to data protection principles. These data, held by controllers who can link internally, must be sufficiently Barnardised or ‘changed’ in order that it is no longer identifiable. DataSHIELD accomplishes this in different ways. The score vectors and information matrices sent to the AC during analysis are the calculations made using the individual-level data; they are not the data themselves. DataSHIELD also uses various statistical data control measures to protect confidentiality, such as was noted earlier, preventing tables returning data including fewer than 5 observations. This form of cell suppression is a recognised method of statistical data control actively used by the UK Government to reduce the possibility of re-identification due to small numbers [27]. For example, the UK Statistics of Trade Act 1947, which regulates the collection of personal information for economic purposes, states in Section 9(5)(a), ‘… no such report, summary or communication [of public information] shall disclose the number of returns received with respect to the production of any article if that number is less than five.’

In addition, there are limits on the different kinds of analysis that can be run on the data, which increases the security of the DataSHIELD process. There is an approved list of the only commands that can be sent through DataSHIELD preventing, as far as is possible, any disclosures of identifiable data which could result from running potentially malicious statistical analyses or through combining a string of other risky commands. We acknowledge that DataSHIELD cannot avoid all disclosures, such
as inferential disclosure [35], as disclosure is a risk inherent to all methodologies.

**Does the Use of Anonymised DataSHIELD Data Respect the Right to Privacy?**

Recent case law has also caused legal scholars to ask whether the disclosure of anonymised data can be seen to breach an individual’s right to privacy, as set down in Article 8 of the UK Human Rights Act 1998 (HRA1998). In *R v. Department of Health, Ex Parte Source Informatics Ltd.* [36], Source Informatics Ltd. sought anonymised data from pharmacists on GP prescribing habits, which they would sell to pharmaceutical companies to use for marketing purposes. The Department of Health argued that anonymisation did not remove the duty of confidence that GPs had with their patients and that patients would not have wanted their data used for such commercial purposes. In the first instance, the court ruled in favour of the Department of Health, stating that the use of the data would go beyond that which was agreed by the participant at the time of prescribing.

[A] breach of confidence occurs necessarily if, without the explicit or implied consent of the confider, the confidant further discloses or uses information obtained in confidence for any purpose other than that for which the information was provided [37].

This judgement was overturned on appeal. One of the grounds for the decision was that there would only be a breach of confidence if, in further disclosing a patient’s data, the pharmacist was taking unfair advantage of the patient. Harm might be caused if the pharmacist acted in a way in which the patient might disapprove and, therefore, cause distress. The Court disagreed and stated that the law in this case was concerned with the protection of the patients’ personal privacy. Patients had no rights to control the further use of the data, provided their privacy was not put at risk. By breaking the link to the patient before disclosure through anonymisation, the patient could no longer be seen as being treated unfairly. However, one might still see a conflict with Article 8.1 of the HRA 1998 which states that, ‘[e]veryone has a right to respect for his private and family life …’ For example, as Beyleveld and Townend [38] argue, if the pharmaceutical companies used anonymised data provided by Source Informatics to develop contraceptive medicines, this might be objectionable to Roman Catholic patients. ‘… [R]endering personal information non-personal is surely not by itself sufficient to preclude a violation of one’s moral integrity.

This case points to the need for data controllers, such as research studies, to ensure that the data they hold are used for purposes within the scientific scope of the study and in concert with consent agreements. As in the HOP study, the ethical approvals and consent materials of each participating study need to be examined to ensure that the use is acceptable [18]. In this way, studies can reassure stakeholders (e.g. participants, ethics committees, funders) that anonymised data will be used in approved ways.

**Can the Processing of Anonymised DataSHIELD Data Still Fall under Data Protection Principles?**

As noted earlier, Recital 26 of the Directive states that anonymised data fall outside data protection principles. However, Beyleveld and Townend [38] have also argued that, using a broad definition, the processing of anonymised data can still be considered processing of personal data because the anonymised data were derived from the personal data. We agree with the position that it would be inappropriate to use anonymisation simply as a means to avoid having to follow data protection principles [38]. The anonymisation in the DataSHIELD process is integral to its design, allowing the approved movement of research data between studies that might have conflicting regulatory schemes or consent provisions, while providing security against unwanted disclosure. However, while no certainty on this point is possible, we believe that we satisfy the need to protect data by ensuring that all direct processing of personal data is conducted within the protected environment of the contributing study; that participating studies have scientific, ethics and data access approvals in place prior to the analysis; that studies are clear that their consents conform with this use of data, and that participants are informed that anonymisation for further disclosure will take place.

**Limitations of This Study and Future Research**

This paper has focused on UK requirements; DataSHIELD is currently being piloted in several multicountry studies, which introduces the need to determine the legal and ethics review requirements for using DataSHIELD in other countries. Work is underway focusing on other international settings. We also acknowledge that DataSHIELD may in the future be used for different purposes, such as quality assurance or for public health surveillance. Work will be needed to investigate any different legislative or ethics review requirements that might be introduced for these uses.
Once studies begin using DataSHIELD, we hope that it will become a recognised and standard methodology. If so, there is a question as to whether ethics approval would be needed for each individual DataSHIELD analysis as opposed to the embedding of DataSHIELD in a biobank or cohort study for use in multiple analyses. We suggest that once approval is given for a study to use DataSHIELD as a methodology, additional approvals for each separate use should not be needed. This is because the DataSHIELD process remains the same regardless of the variables chosen; only the research questions and variables change. Approval for new research investigations should be confirmed or obtained, but not a new approval to use DataSHIELD in an analysis. There is precedent for our position. Each use of the CRIP database was deemed not to require separate ethics approval [7]. However, there are significant differences between DataSHIELD and CRIP, and, therefore, this aspect would need to be explored with representatives of ethics committees.

Is additional consent needed from participants for the use of their anonymised data in medical research? We agree with others that such a requirement could seriously hamper epidemiological research [see 29]. There is a recognised need, and support from the UK government, for the sharing of anonymised data for medical research [39]. However, critics argue that anonymisation does not in fact ensure privacy [40]. This position is supported by the increasing examples of re-identification of anonymised data through varying methodologies using publically available databases of information [9–11]. Addressing this possibility is an important consideration for any research study promising to protect confidential data. DataSHIELD seeks to answer this by leaving individual-level data within the protection of the contributing studies and only sharing summary statistics that, even when combined with the personal data, could not identify any one individual. However, we also acknowledge that continued efforts will need to be made to ensure we are using best practice methods to protect confidentiality and restrict disclosure.

Conclusion

Collaborative studies are increasingly using the power of large datasets to study complex scientific questions. Sharing these resources is vital [41] but must be done in accordance with recognised ethico-legal practices. We recognise that there are concerns about anonymisation as a ‘cure all’ when seeking to protect the confidentiality of participants and enable sharing. However, we strongly support the position that, ‘[f]ar from potentially harming participants and researchers, methods and emerging technologies that work within a regulatory framework or legislation demonstrate how anonymity may facilitate innumerable benefits.’ [42]. We believe DataSHIELD is one of these technologies that can in equal measures respect the wishes of participants to protect their data while enabling those data to be shared for further scientific discovery.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 261433.
DataSHIELD and UK Law

DOI: 10.1159/000360255

13-920-government-response-administration-of-data-protection/the-data-protection-act.pdf.

23 Ihaka R, Gentleman R: R: a language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.

24 Romeo-Casabona CM, Nicolas P, Knoppers BM, Joly Y, Wallace SE, Chalmers D, Dyke S, Kennedy K, Troncoso A, Kaan T, Rial-Sebbag E: Legal aspects of genetic databases for international biomedical research: the example of the International Cancer Genome Consortium (ICGC). Law Hum Genom Rev 2012;37:15–34.

25 International Epidemiology Association: Guidelines for proper conduct in epidemiologic research. 2007. http://ieaweb.org/good-epidemiological-practice-gep/.

26 Information Commissioner: Anonymisation: managing data protection risk code of practice. Wilmslow, Information Commissioner’s Office, 2012. http://ico.org.uk/for_organisations/guidance_index/-~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf.

27 Data protection. 2013. https://www.gov.uk/data-protection/the-data-protection-act.

28 Joly Y, Dove ES, Knoppers BM, Bobrow M, Chalmers D: Sharing data in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). PLoS Comput Biol 2012;8:e1002549.

29 Miola J: Owning information – anonymity, confidentiality and human rights. Clin Ethics 2008;3:116–120.

30 Hordern V: Does statistical data equate to anonymised data: Debate. Data Protec Law Pol 2011:12–13.

31 Common Services Agency v Scottish Information Commissioner [2008] UKHL 47.

32 Barnard GA: Statistical inference. J R Stat Soc 1949;11:115–149.

33 Doyle P, Lane JI, Theeuwes JJM, Zayatz LV (eds): Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam, Elsevier, 2001.

34 R (Department of Health) v Information Commissioner [2011] EWHC 1430 (Admin).

35 Duncan GT, Jabine TB, de Wolf VA (eds): Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Washington DC, National Academies Press, 1993.

36 R v Department of Health ex parte Source Informatics Ltd. [2001] 1 All ER 786.

37 Belsey D, Histed E: Betrayal of confidence in the court of appeal. Med Law Int 2000;4:277–311.

38 Belsey D, Townend DMR: When is personal data rendered anonymous? Interpreting Recital 26 of Directive 95/46/EC. Med Law Int 2004;6:73–86.

39 Improving access for research and policy: The Government response to the report of the Administrative Data Taskforce. 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206873/bis-13-920-government-response-administrative-data-taskforce.pdf.

40 Ohm P: Broken promised of privacy: responding to the surprising failure of anonymization. UCLA Law Rev 2010;57:1701–1777.

41 Creating a Global Alliance to enable responsible sharing of genomic and clinical data. 2013. http://www.broadinstitute.org/files/news/pdfs/GAWWhitePaperJune3.pdf.

42 Knoppers BM, Dove E, Litton JE, Niefeld J: Questioning the limits of genomic privacy. Am J Hum Genet 2012;91:577–578.