Anonymising and sharing individual patient data

Khaled El Emam,[1, 2] Sam Rodgers,[3] Bradley Malin[4]

[1]Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada

[2]Faculty of Medicine and School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa

[3]Earls Court Health and Wellbeing Centre, London, UK

[4]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA

Correspondence to: K El Eman kelemam@ehealthinformation.ca

Cite this as: *BMJ* 2015;350:h1139
doi: 10.1136/bmj.h1139

There is a strong movement to share individual patient data for secondary purposes, particularly for research. A major obstacle to broad data sharing has been the concern for patient privacy. One of the methods for protecting the privacy of patients in accordance with privacy laws and regulations is to anonymise the data before it is shared. This article describes the key concepts and principles for anonymising health data while ensuring it remains suitable for meaningful analysis.

There is increasing pressure to share individual patient data for secondary purposes such as research.[1–3] For example, research funding agencies are strongly encouraging recipients of funds to share data collected by their projects.[4–6] The expected benefits from sharing individual patient data for health research purposes include: it ensures accountability in results and that reported study results are valid, it allows researchers to build on the work of others more efficiently and to perform individual patient data meta-analyses to summarise evidence, and it decreases the burden on research subjects through the reuse of existing data.[7] In many instances, however, patient privacy concerns have been perceived as a key barrier for making individual patient data available.[3 8]

There are two legal mechanisms that would permit data custodians to share patient data for secondary purposes (unless there is an exemption in the law): (*a*) consent and (*b*) anonymisation. If the data was originally collected in a medical context, then consent for unanticipated secondary analyses is often not obtained in advance. It is not always practical to go back and obtain consent from a large number of patients, and there is evidence of systematic consent bias whereby consenters and non-consenters differ on important characteristics.[9–11] As a consequence, it is challenging to rely on consent as the primary mechanism for sharing data. With respect to the second option, there is evidence that many research ethics boards will permit the sharing of patient data without consent for research purposes if it is anonymised.[12] (The term "de-identification" is more commonly used in North America while "anonymisation" is more commonly used in Europe; for this article, we treat the terms as equivalent.)

Many jurisdictions, including those in North America and Europe, do not designate anonymised health data as personal information.[7] Therefore, such data would no longer be covered by privacy laws, allowing it to be used and disclosed for any secondary purpose. However, there is an expectation that the anonymised data will be used only for purposes that are legitimate, in a manner that would not surprise the patients, and not in a discriminatory or stigmatising manner. This expectation has been made explicit in the EU context,[13] and falls under a privacy ethics framework outside the European Union.[14]

When sharing patient data for secondary purposes it is important to be mindful of patient trust. While patients are supportive of the use of their data for research,[7] often there is an expectation that that data will be adequately anonymised. Trust is important because there is evidence that patients adopt privacy protective behaviors, such as lying and not seeking care, when they have concerns about how their health information may be shared.[15]

Definitions of anonymity in privacy laws and regulations do not provide an operational method to follow for anonymising health information. Even the concept of anonymous or non-identifiable data is ambiguous. For example, the European Data Protection Directive 95/46/EC states that "'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity"; and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule of 1996 in the US notes that "Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information." This ambiguity contributes to heterogeneity and inconsistency in actual anonymisation practices for health data.

The subdiscipline of statistics known as disclosure control has developed a substantial body of knowledge around anonymisation techniques.[16 17] In this article we describe the key concepts and principles behind the anonymisation of health data in an effort to find a common language and mitigate current inconsistencies. As a running example, we will use the Ontario (Canada) birth registry dataset (known as BORN) to illustrate various points. BORN is a population registry of all births in the province. The data is collected from hospitals, clinics, midwives, and the provincial newborn screening laboratory and stored in a data warehouse. The data is then used and disclosed for research and public health purposes.[18]

Definitions

From a technical perspective, ensuring anonymity equates to ensuring that the probability of assigning a correct identity to a record in a dataset is very small. This probability can be conditional on other factors, such as the skills required and resources available to an adversary seeking to re-identify a record.[7] When data is shared, it is not possible to ensure that the probability of re-identification is zero, but it is possible to ensure that the probability is very small.

Existing standards and guidelines tend to divide the variables in a dataset into two groups: direct identifiers and quasi-identifiers. The direct identifiers are features that permit direct recognition or communication with the corresponding individuals, such as personal names, email addresses, telephone numbers, and social insurance numbers. Quasi-identifiers are features that can indirectly identify individuals, such as their date of birth, death, or clinic visit, residence postal code, and ethnicity. Quasi-identifiers include demographics and socioeconomic information. Both types of variables must be addressed during anonymisation.

In the case of the BORN registry, variables such as the mother's name and health insurance number are designated as direct identifiers. These variables are removed before data arrives at the registry. Sometimes there are unique identifiers that need to be retained to allow linking of all of the records that belong to the same mother (for example, to track multiple births) such as a medical record number. Because a medical record number is often considered a patient identifier as well, it is converted to a pseudonym. The data is then called "pseudonymised." Pseudonymous data is still considered personal information under the European Data Protection Directive 95/46/EC[19] and should not be treated as anonymous.

To date, all known successful re-identification attacks (excluding genetic data) were performed on pseudonymous data.[20] Adversaries performing such an attack attempt to determine the identity of individuals in a dataset that has been shared. Known re-identification attacks are performed almost exclusively by researchers and the media.[20]

The motives of the media are believed to be to show that shared data is unsafe (which makes for a good story) or to contact individuals and their families for a story. Academics perform these attacks to publish new computational algorithms for attacking databases and also to show weaknesses in available databases. In general, such "white hats" get recognition for finding weaknesses in systems and databases. We consider two examples below.

An example of a media initiated re-identification attack is when a national Canadian broadcaster re-identified an individual in the adverse drug event database from Health Canada. The purpose was to report on the adverse events associated with a drug, and they wanted to interview the family of the deceased individual who was re-identified.[21] The re-identification attack used publicly available obituaries to match on age, location, and date of death to determine the identity of the 26 year old woman who had died while taking the drug in question.

A recent example of a successful re-identification attack by a team of a reporter and an academic was performed on a hospital discharge database. The department of health in Washington state in the United States was sharing pseudonymised data with few restrictions on who could access the data and what the data recipient could do with it. In this attack, the adversaries used information from newspaper articles about vehicle accidents and reports involving hospitalisations of famous people in the media to re-identify individuals in the hospital discharge database.[22 23] This was accomplished by combining the discharge data with publicly available phone number directories and voter registration lists. Specifically, in this attack the adversary leveraged knowledge about the date of admission, the injury code, the age of the patient, which hospital was visited, the ZIP code of the patient, whether it was a weekend admission, as well as the gender and race of the patient. This amounted to 11 quasi-identifiers that were leveraged to attack the database.

In both of the above cases the successful re-identification attack used quasi-identifiers. It is therefore important to protect the quasi-identifiers as well as the direct identifiers.

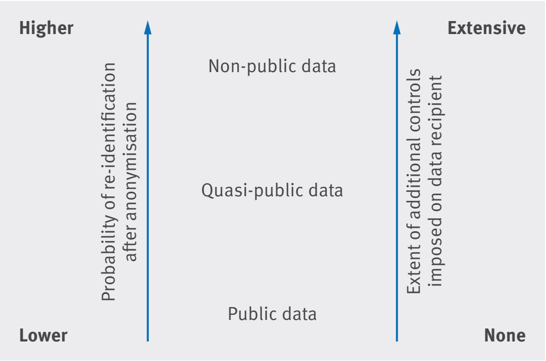
Types of data sharing

There are three general ways to share data for secondary purposes: public, quasi-public, and non-public.

Public data has the least amount of restrictions placed on it. Such public data is available, typically online, for anyone to download either free or for a nominal fee. Many national statistical agencies release census and national survey data as public data. Some of this survey data includes health information. There are also publicly available clinical trials data from the International Stroke Trial[24] and data posted in the Dryad online open access data repository.[25 26]

Quasi-public data has additional restrictions imposed on it in the form of a "terms of use." This is a contract that the data recipient signs (or clicks through if it is online). The terms of use often includes a prohibition on attempting to re-identify the data, contacting any of the patients, linking the data with other datasets, and sharing the data with any third party. Also, all data recipients must register so that their identity is known to the data custodian. Data competitions serve as illustrative examples of the use of quasi-public data. The Heritage Health Prize, for instance, was a $3 000 000 competition whereby the winner built a predictive model for readmissions using a quasi-public dataset from the Heritage Provider Network.[27] The 2013 Cajun Code Fest[28] was another $25 000 competition where the entrants built software applications that used quasi-public data to support decision making for patients and providers. In both competitions, all entrants had to register and agree to the terms of use contract before they were allowed to access the data.

Non-public data has the most restrictions placed on it. In this case the data recipient would need to sign a full contract that, in addition to the above specifications, includes a prescriptive set of security and privacy


Higher
Lower
Probability of re-identification after anonymisation
Non-public data
Quasi-public data
Public data
Extent of additional controls imposed on data recipient
Extensive
None

Extent of anonymisation that needs to be applied for different types of data releases balanced against other controls

controls that the data recipient needs to have in place, such as encrypting their computers and providing privacy training to the analysts who will work with the data. The data custodian may also reserve the right to audit recipients to ensure that they comply with all of the conditions.

The data needs to be anonymised in all of the three cases above. However, the acceptable probability of re-identification would vary. For a public data release the probability needs to be quite low because there are no other controls that can be put in place. However, for non-public data a higher probability would be acceptable because other security, privacy, and contractual controls would be put in place. This balancing of controls to manage the risk is illustrated in the figure.

The above distinctions mean that the same data can be sufficiently anonymised in different ways depending on the context of the data release. Accounting for the context of the data release when deciding on how to anonymise is consistent with existing best practices and regulatory guidance.[29–31]

The mechanism of data release can also vary. For example, individual patient data may be provided to a researcher for download, or the researcher may get access to the individual patient data through a portal that does not allow any data to be downloaded. In the latter case all of the analysis must happen on the portal itself. Some data custodians require the researcher to be physically present in a secure room in order to access individual patient data. Each of these mechanisms has a different set of controls imposed on the researcher, and therefore the acceptable probability of re-identification would be set accordingly.

Table 1 | Probability of re-identification of anonymised data in BORN (Ontario birth registry dataset) for various combinations of quasi-identifiers

Mother's date of birth	Baby's date of birth	Mother's postal code	Baby's sex	Probability of re-identification*
X				0.014
	X			0.005
X	X			0.88
X	X	X		1.00
X	X	X	X	1.00
X	X		X	0.91
	X	X		0.98
X		X		0.85
		X		0.19

X indicates that a variable is included in the calculation of probability.
*Probability was measured using the average re-identification risk metric defined elsewhere.[7]

Measuring the probability of re-identification

The balancing described above is premised on the ability to measure the probability of re-identification. Several metrics have been developed for measuring the probability of re-identification.[7] These can be applied for datasets over a large population or for samples derived from the population. The BORN registry is an example of a population dataset because it includes all births in Ontario. In that case, the probability of re-identification can be directly measured from the data. A sample dataset could be, for example, a clinical trial with diabetic patients (because only a subset of all patients with diabetes will participate in that trial). In the case of the clinical trial dataset, the probability of re-identification would have to be estimated from the data.

To start with, the probability of re-identification will depend on two factors: (*a*) which quasi-identifiers are included in the shared dataset and (*b*) the extent to which the data has been perturbed (or modified).

In the BORN registry, variables such as the baby's date of birth and sex and the mother's date of birth and postal code are designated quasi-identifiers. They could also be discovered by an adversary for various reasons: births are commonly announced, residence information is available from sources such as the Whitepages (Canadian and US telephone and address directories), and basic demographics are generally available from a variety of public resources.[32] We can illustrate how the probability of re-identification is affected by the selected quasi-identifiers.

Table 1 shows the probability of re-identification for different combinations of quasi-identifiers in BORN. The dataset we use has 919 710 births from 2005 to 2011. This probability will vary depending on which quasi-identifiers are included in the released data. In general, the more quasi-identifiers that are included in the released data, the greater the probability of re-identification. Some quasi-identifiers have a substantial impact, such as the Canadian six-digit postal code, followed by the mother's date of birth, whereas other quasi-identifiers have little to no impact (such as the baby's sex). The inclusion of all four quasi-identifiers leads to a high probability of re-identification because at that level of detail almost all births are unique.

Data transformations and data quality

If the probability of re-identification is deemed to be too high, then various perturbation techniques can be applied to reduce it.[14] For example, if all quasi-identifiers in table 1 need to be shared without perturbation, it is almost certain that re-identification can happen.

One of the simplest ways to perturb the data is to reduce the precision of data fields through generalisation. This approach is used quite often in practice. As an illustration, it is natural for a date of birth to be generalised into

Table 2 | Changes in probability of re-identification of anonymised data in BORN (Ontario birth registry dataset) for different levels of generalisation of quasi-identifiers

Scenario	Mother's date of birth or age	Baby's date of birth	Mother's postal code	Baby's sex	Probability of re-identification*
S1	Year	day, month, year	3 character	Unchanged	0.973
S2	Year	month, year	3 character	Unchanged	0.677
S3	Age in 5-year groups	month, year	3 character	Unchanged	0.327
S4	Age ≤19, 20–30, 30–40, >40	month, year	3 character	Unchanged	0.23
S5	Age ≤19, 20–30, 30–40, >40	month, year	1 character	Unchanged	0.007
S6	Year	month, year	1 character	Unchanged	0.034
S7	Age in 5-year groups	quarter, year	3 character	Unchanged	0.152
S8	Age ≤19, 20–30, 30–40, >40	quarter, year	3 character	Unchanged	0.1

*Probability was measured using the average re-identification risk metric defined elsewhere.[7]

a month and year of birth. Generalisation is, in many instances, considered to be an acceptable strategy for protection because it is consistent with how the data will be analysed. For example, if the analysis only requires the year of birth of the mother, then generalising the mother's date of birth in BORN will reduce the probability of re-identification and will be consistent with the intended analysis.

Table 2 depicts the probability of re-identification after various generalisations were applied to the BORN quasi-identifiers. Simple changes to the data can result in substantial reductions in the probability of re-identification. Which generalisation should be chosen is determined using a combination of two methods: (*a*) a data analyst subjectively judges whether a particular generalisation would affect the ability to analyse the data, and (*b*) formal metrics are applied to evaluate data utility, such as the entropy in the resulting records.[10]

In table 2, scenario S1 reduces the precision of the mother's date of birth to a year and the postal code to the first three characters, but the probability of re-identification remains quite high. By contrast, scenarios S5 and S6 have the lowest probability of re-identification, but the postal code is truncated to the first character only. This precludes most meaningful geospatial analysis. The lowest probability that maintains location information is reached with scenario S8, with the baby's date of birth converted to quarter and year and the mother's age is categorised as ≤19, 20–30, 30–40, or >40 years. However, the changes in S8 reduce the utility of the data because details around the exact age of the infant at certain time points cannot be calculated, and geospatial analysis is still limited by the three character postal code.

ANONYMISATION FAQS

- Is it necessary to obtain patient consent to anonymise health data or to share anonymised data?

  In most jurisdictions, including the European Union, anonymisation is considered a permitted use.[13] This means that it is not necessary to obtain patient consent to anonymise the data.
- Can data on rare diseases be anonymised?

  The presence of a rare disease does not necessarily make it impossible to anonymise. If the dataset is a sample from the population of patients with that disease, then the probability of re-identification may still be small. If the rare disease is not visible then that reduces the likelihood that an adversary would know that someone has that disease.[37]
- Will advances in technology and the greater availability of data increase the risk of re-identification?

  Anonymisation is typicaly time limited to account for changes in technology and the availability of other data that can be used to re-identify individuals. This time limit is typically 18–24 months. After that time has elapsed, the risk of re-identification needs to be re-evaluated to determine if circumstances make the originally anonymised data high risk. This is possible to achieve for non-public datasets where permission to use a dataset is time limited and the data use agreement stipulates a re-assessment of re-identification risk. For public data, the initial anonymisation needs to be more stringent to be applicable for a longer period since it is not possible to "call back" a public dataset.

Better methods of perturbation can be used than simple generalisation. These computational methods can reduce the amount of distortion to the data (such as allowing more granularity than the three character postal code) and produce higher data quality.[14,33]

In practice, when there are many quasi-identifiers in a dataset, simple techniques such as generalising the values for all the records in the same way are unlikely to produce datasets that are analytically useful. With just the four quasi-identifiers in Table 2, the acceptable generalisations were already approaching the limits of data utility. However, as mentioned earlier, recent re-identification attacks leveraged as many as 11 quasi-identifiers.[22,23] To maintain the utility of the data, more sophisticated methods can be applied that retain details in dates and geospatial information during the anonymisation process.[14]

When to stop

A practical question that the data custodian needs to answer is how much generalisation is enough? For instance, are all the solutions in table 2 that are below a probability of re-identification of 0.2 acceptable from a risk perspective? There are precedents (regulatory, legal, and practical) going back decades for what is an acceptable probability of re-identification for public and non-public data releases.[7] These precedents provide a range of possible acceptable thresholds that can justifiably be used. In general, they vary from an acceptable probability of 0.33 to 0.05.[7]

There are instances where anonymisation schemes do not include risk measurement nor the setting of thresholds to ensure that the probability of re-identification is acceptable.[34-36] For example, these schemes provide a fixed list of quasi-identifiers that should be removed from the dataset. These approaches cannot

ANONYMISING CLINICAL TRIALS DATA

Regulators such as the European Medicines Agency are planning to make data from clinical trials more generally available.[38 39] Initially, the contents of clinical study reports will be made available under a two-track process, with broad public access through a portal and the ability to download for a narrower set of identified users.[40] In a second phase, individual patient data will be made available. However, the agency cannot collect individual patient data only for the purpose of sharing it and needs to formulate policies on how to use the individual patient data for scientific review as well. This has resulted in some delays in formulating a policy for sharing individual patient data.

In anticipation of individual patient data being made available by regulators, or the requirement by them to do so, manufacturers have already started putting in place policies and infrastructure for sharing individual patient data.[41] Recent examples include:

- The GlaxoSmithKline trials repository,[42 43] which now has multiple pharmaceutical companies using it to manage the data request process and share data (www.clinicalstudydatarequest.com)
- The Data Sphere project, a consortium of pharmaceutical companies, sharing data from the control arm of oncology trials[44 45]
- The Yale University Open Data Access project, which is initially making trial data from Medtronic available[46 47]
- The Immport Immunology Database and Analysis Portal[48]

Furthermore, some pharmaceutical companies are creating their own company-specific portals to facilitate the sharing of their own datasets, and these are typically accessible through their corporate websites.

Given that trial participants are often from multiple sites across the world, anonymisation practices for the data must meet the regulatory requirements globally. This means that the burden of evidence that the probability of re-identification is acceptably small is not trivial because regulators in different jurisdictions do not use the same standards. Organisations such as the European Medicines Agency could help address such gaps by providing or recommending robust and scalable methods that can provide quantitative anonymity assurances while producing high quality data.

provide assurance that the probability of re-identification is small for any single dataset because the actual quasi-identitifers may differ from the list. Moreover, their application may result in datasets being excessively perturbed. Therefore, such approaches would not be appropriate for complex datasets. Knowing when to stop perturbing the data is important to balance privacy protection and data utility.

Conclusions

Methods for measuring the risk of re-identification can be used to decide how much to anonymise health data for different types of data release. Perturbation that retains sufficient data quality requires data-centric methods rather than simplistic rules regarding how to generalise fields. Anonymisation methods cannot ensure that the risk of re-identification is zero, but this is not the threshold that is expected by privacy laws and regulations in any jurisdiction. Strong precedents exist for choosing suitable probability thresholds for anonymising data. There is a need for anonymisation standards that can provide operational guidance to data custodians and promote consistency in the applications of anonymisation.

Contributors: KEE's research over the past 10 years has been focused on privacy issues related to electronic health information, particularly the policy and technological obstacles and solutions to data anonymisation and the sharing of health data for secondary purposes. He has published three books on this topic covering policy, legal, and methodology issues, as well as a series of case studies to show how to anonymise health data. SR was formerly the clinical chief information officer for NHS Central London CCG. He was the clinical lead on implementing a single primary care IT system for Central London CCG and contributed to the development of the data storage and analysis platform which will house patient data in the North West London area. BM conducts research on methods and software tools to anonymise health data for secondary purposes. He contributed to the development of the US HIPAA de-identification guidelines.

Funding: BM's work on this article was funded by grants R01LM009989 (National Library or Medicine, National Institutes of Health) and U01HG006385 (National Human Genome Research Institute, National Institutes of Health).

Competing interests: The authors have read and understood the BMJ Group policy on declaration of interests and declare the following interests: KEE and BM have financial interests in Privacy Analytics, a University of Ottawa and Children's Hospital of Eastern Ontario spin-off company which develops anonymisation software for the health sector.

Ethical approval: The analysis of the BORN registry data described in this article was approved by the Research Ethics Board of the Children's Hospital of Eastern Ontario Research Institute.

1 Gøtzsche PC. Why we need easy access to all data from all clinical trials and how to accomplish it. *Trials* 2011;12:249.

2 P. Vallance, I. Chalmers. Secure use of individual patient data from clinical trials. *Lancet* 2013;382:1073–4.

3 Olson S, Downey AS. *Sharing clinical research data: workshop summary*. National Academies Press, 2013.

4 Medical Research Council. Research policy & ethics: Data sharing. 2015. www.mrc.ac.uk/research/research-policy-ethics/data-sharing/.

5 National Institutes of Health. Final NIH statement on sharing research data. 2003. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

6 Wellcome Trust. Sharing research data to improve public health: full joint statement by funders of health research. 2011. www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm.

7 El Emam K. *Guide to the de-identification of personal health information*. CRC Press, 2013.

8 Castellani J. Are clinical trial data shared sufficiently today? Yes. *BMJ* 2013;347:f1881.

9 Kho M, Duffett M, Willison D, Cook D, Brouwers M. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* 2009;338:b866.

10 El Emam K, Dankar F, Issa R, Jonker E, Amyot D, Cogo E, et al. A globally optimal k-anonymity method for the de-identification of health data. *J Am Med Inform Assoc* 2009;16:670–82.

11 El Emam K, Jonker E, Moher E, Arbuckle L. A review of evidence on consent bias in research. *Am J Bioethics* 2013;13:42–4.

12 Willison DJ, Emerson C, Szala-Meneok KV, Gibson E, Schwartz L, Weisbaum KM, et al. Access to medical records for research purposes: varying perceptions across research ethics boards. *J Med Ethics* 2008;34:308–14.

13 El Emam K, Alvarez C. A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. *Int Data Privacy Law* 2015;5:73–87.

14 El Emam K, Arbuckle L. *Anonymizing health data: case studies and methods to get you started*. O'Reilly, 2013.

15 Malin BA, Emam KE, O'Keefe CM. Biomedical data privacy: problems, perspectives, and recent advances. *J Am Med Inform Assoc* 2013;20:2–6.

16 Willenborg L, de Waal T. *Statistical disclosure control in practice*. Springer-Verlag, 1996.

17 Duncan G, Elliot M, Salazar G. *Statistical confidentiality: principles and practice*. Springer, 2011.

18 BORN Ontario. https://www.bornontario.ca/.

19 Article 29 Data Protection Working Party. Opinion 05/2014 on anonymization techniques (WP216). 2014. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

20 El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. *PLoS One* 2011;6:e28071.

21 El Emam K, Dankar FK, Neisa A, Jonker E. Evaluating the risk of patient re-identification from adverse drug event reports. *BMC Med Inform Decis Mak* 2013;13:114.

22 Sweeney L. Matching known patients to health records in Washington State data. Harvard University, Data Privacy Lab, 2013.

23 Robertson J. States' hospital data for sale puts privacy in jeopardy. *Bloomberg News* 2013 Jun 5. www.bloomberg.com/news/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy.html.

24 Sandercock PA, Niewada M, Członkowska A, the International Stroke Trial Collaborative Group. The International Stroke Trial database. *Trials* 2011;12:101.

25 Dryad Digital Repository. http://datadryad.org/.

26 Haggie E. PLOS Genetics partners with Dryad. *PLOS Biologue* 2013. http://blogs.plos.org/biologue/2013/09/18/plos-genetics-partners-with-dryad/.

27 El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, et al. De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. *J Med Internet Res* 2012;14:e33.

28 Cajun Code Fest. 2013. http://cajuncodefest.org/.

29 Anonymisation: managing data protection risk code of practice. Information Commissioner's Office, 2012. https://ico.org.uk/media/1061/anonymisation-code.pdf.

30 Health System Use Technical Advisory Committee, Data De-identification Working Group. *'Best practice' guidelines for managing the disclosure of de-identified health information*. Canadian Institute for Health Information, 2010.

31 US Department of Health & Human Services. *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule*. Department of Health and Human Services, 2012.

32 El Emam K, Jonker E, Sams S, Neri E, Corporation T, Neisa A, et al. *Pan-Canadian de-identification guidelines for personal health information*. Office of the Privacy Commissioner of Canada, 2007.

33 Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J Biomed Inform* 2014;50:4–19.

34 Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials* 2010;11:9.

35 Shostak J. *De-identification of clinical trials data demystified*. SAS Users Group, 2006.

36 Hughes S, Wells K, McSorley P, Freeman A. Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach. *Pharmaceut Statist* 2014;13:179–83.

37 Eguale T, Bartlett G, Tamblyn R. Rare visible disorders/diseases as individually identifiable health information. *AMIA Annu Symp Proc* 2005:947.

38 European Medicines Agency. Release of data from clinical trials. 2013. www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/general/general_content_000555.jsp&mid=WC0b01ac0580607bfa.

39 Eichler H-G, Abadie E, Breckenridge A, Leufkens H, Rasi G. Open clinical trial data for all? A view from regulators. *PLoS Med* 2012;9:e1001202.

40 European Medicines Agency. European Medicines Agency policy on publication of data for medicinal products for human use. 2014. www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.

41 PhRMA, EFPIA. Principles for responsible clinical trial data sharing. 2013. www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.

42 Nisen P, Rockhold F. Access to patient-level data from GlaxoSmithKline clinical trials. *N Engl J Med* 2013;369:475–8.

43 Harrison C. GlaxoSmithKline opens the door on clinical data sharing. *Nat Rev Drug Discov* 2012;11:891–2.

44 Hede K. Project data sphere to make cancer clinical trial data publicly available. *J Natl Cancer Inst* 2013;105:1159–60.

45 Bhattacharjee Y. Pharma firms push for sharing of cancer trial data. *Science* 2012;338:29.

46 Krumholz HM, Ross JS. A model for dissemination and independent analysis of industry data. *JAMA* 2011;306:1593–4.

47 Yale School of Medicine. Center for Outcomes Research & Evaluation (CORE): YODA Project. http://medicine.yale.edu/core/projects/yodap.

48 ImmPort: bioinformatics for the future of immunology. https://immport.niaid.nih.gov/.