Chapter 14
Key Law and Policy Considerations for Clinical Bioinformaticians

Mark Phillips

Abstract This chapter describes five key areas in which clinical bioinformatics activities are regulated by law and policy. These are, namely, open-data requirements, consent practices, anonymization strategies, restrictions on cross-border data transfer, and prohibitions on genetic discrimination. The discussion draws on examples of norms that are currently in effect in North America, Europe, Asia, and Oceania in order to illustrate the ways in which positions on various specific questions can either mutually converge or deviate from one another around the globe. The tension that animates virtually all of the debates throughout this area, whether explicitly or through proxy issues, this chapter argues, is between the promotion of the interests of research participants—particularly in ensuring data privacy—on the one hand, and in establishing a landscape optimized to best promote medical research discoveries, on the other.

Keywords ELSI • Privacy • Data protection • Open data • Funding agency policy

14.1 Introduction

Clinical bioinformaticians are governed by many of the same legal and policy norms as are researchers in genomics and the broader -omics fields. Physicians and health care practitioners who provide personalized medicine to their patients additionally remain bound by their professional ethics duties and laws applicable to the provision of health care.

Two new trends add complexity. First, a flurry of new rules have been adopted in recent years by law and policymakers who are scrambling to address ethical and privacy concerns that have emerged—and that often remain poorly understood—as a direct result of the rapid development of genomic technologies. Second, the increasing prevalence of data sharing, often across borders, as well as outsourcing to cloud service providers can mean that health projects must simultaneously contend with rule sets in multiple jurisdictions. Disparities between the rule sets

M. Phillips, B.Sc., LL.B., B.C.L. (✉)
Centre for Genomics and Policy, McGill University, Montreal, 740, avenue Dr. Penfield, suite 5200, Montreal, QC H3A 0G1, Canada
e-mail: mark.phillips2@mcgill.ca

X. Wang et al. (eds.), *Application of Clinical Bioinformatics*,
Translational Bioinformatics 11, DOI 10.1007/978-94-017-7543-4_14

can give rise to incompatibilities that may even needlessly make some medical projects impractical.

Despite this fragmentation, researchers have initiated significant efforts aimed at harmonizing these obligations at national, regional, and international levels. Although these efforts remain in the preliminary stages, their emphasis provides a helpful window through which to frame a general discussion of the law and policy as it currently affects clinical bioinformatics. Each of this chapter's following five sections describe one key area around which these discussions have centred, specifically (Sect. 14.2) open-data requirements; (Sect. 14.3) consent practices; (Sect. 14.4) anonymization strategies; (Sect. 14.5) restrictions on cross-border data transfer; and (Sect. 14.6) prohibitions on genetic discrimination. Each section draws on examples from existing laws and policies.

Before turning to these topics, the following subsections briefly distinguish and explain the two main sources of the duties that will be discussed: data-privacy law and funding agency policy.

14.1.1 Data-Privacy Law

Some countries use the term *privacy law* to describe law aimed at protection of personal data, while others speak of *data protection law*. This chapter follows the emerging trend of using the term *data-privacy law* to encompass both sets of laws.

Laws can guarantee the right to data privacy at the highest level of legal norms: in constitutions. The European Union (EU) has long been at the forefront of data-privacy law development, and its *Charter of Fundamental Rights of the European Union*—although not itself a formal constitution—conceives of personal data protection as a freestanding, fundamental right held by everyone (Fig. 14.1).

Across the Atlantic, the words 'data protection' or 'privacy' appear neither in the *United States Constitution* nor the *Canadian Charter of Rights and Freedoms*. But Supreme Court decisions have established that a degree of constitutional privacy protection is a necessary accessory to other explicit constitutional rights. The most prominent example is that the constitutional right to be secure against unreasonable searches and seizures has been found to necessarily flow from an underlying assumption that people enjoy a reasonable right to privacy (*Hunter v. Southam* 1984; *Katz v. United States* 1967).

The vast majority of data privacy law norms that are relevant in everyday practice are described in regular statutes. Data-privacy law varies significantly between countries: both in terms of the degree of protection provided, as well as the overall framework in which they are set out.

For decades, the European Union (EU) has encouraged its member countries to achieve a measure of harmonization. The preeminent EU data-privacy vehicle is currently the *Data Protection Directive 95/46/EC* (EU Directive), now 20 years old, which requires that each state subject to the *Directive* enact data-privacy legal protections that meet the minimum standards it describes. The EU Directive

1. Everyone has the right to the protection of personal data concerning him or her.
2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
3. Compliance with these rules shall be subject to control by an independent authority.

Fig. 14.1 Article 8 of the *Charter of Fundamental Rights of the European Union*

maintains flexibility, however, by allowing each country a degree of leeway in their preferred implementation of the rules, to account for the legal traditions and social context in each country.

The EU Directive is now set to be superseded by a new *General Data Protection Regulation* (EU Regulation) in 2016. The new EU Regulation will further harmonize existing rules, as unlike the EU Directive, its rules will be directly enforceable throughout all of the European Economic Area and EU member states. At the time that this chapter was written, despite numerous draft iterations that had appeared, the text of the EU Regulation had not yet been finalized.

In contrast with this unified approach in the EU, the United States has relied on both a patchwork of highly specific laws and policies that address privacy, which exist alongside industry self-regulation mechanisms. Several federal laws include provisions that bear on privacy and that may be of particular interest to clinical bioinformaticians. These include the *Health Insurance Portability and Accountability Act of 1996* (HIPAA); the *Federal Policy for the Protection of Human Subjects* (Common Rule); and the *Genetic Information Nondiscrimination Act of 2008* (GINA).[1]

The siloed US approach has the advantage that each narrow topic was given lawmakers' undivided attention as the rules were drawn up, but it also comes with significant shortcomings. The first problem likely to be faced in practice is the difficulty in identifying which of the numerous federal and state laws do or do not regulate any given entity. HIPAA applies only to a list of covered entities including healthcare providers, health plans, and healthcare clearinghouses, as well as the covered entities' business associates and their subcontractors. The Common Rule applies to most organizations receiving federal funding for research. GINA applies to insurance companies and employers, and prohibits certain forms of discrimination based on genetic information.

But these laws can also apply to clinical bionformaticians in ways that may not be initially obvious. When a researcher receives genetic data from a HIPAA-covered entity for use in health research, for example, that researcher is bound to conform to the HIPAA Privacy Rule. As for the Common Rule, in one case in 2010, despite that it was unclear whether the Common Rule was legally enforceable against direct-to-consumer genetic testing company 23andMe, the company's non-compliance with the law's provisions was nonetheless argued to be a valid consideration in determining whether research based on the company's data was fit

[1] Each of these laws have been significantly amended since they were initially enacted.

for academic publication (Tobin et al. 2010). This last anecdote is one illustration of clinical bioinformaticians' interest in complying with generally accepted research and data-stewardship best practices even when the relevant law or policy sets a lower standard. This is especially true when the projects will persist over time.

A second shortcoming associated with the piecemeal US approach to data-privacy law is the risk of unintended gaps in protection. For example, although GINA provides legal protection against discrimination on the basis of asymptomatic genetic features, and the *Americans with Disabilities Act of 1990* protects against discrimination on the basis of disease that has manifested itself and imposes a substantial limitation, experts have cautioned that it remains unclear whether US law provides any protection against discrimination in situations between those two extremes, such as discrimination on the basis of mild symptoms (Rothstein 2008) or discrimination based on predictions made by machine learning techniques on genetic data (Horvitz and Mulligan 2015).

14.1.2 Funding Agency Policies

In addition to being subject to data-privacy law, clinical bioinformaticians must also often subject themselves to institutional policies, often as a condition of receiving funding. While data-privacy laws are enforceable either by national courts or by administrative entities to which the law delegates this task, the direct penalty for failing to adhere to funding policies is aimed at the professional medical activity itself. Funding may be withdrawn from a given project, and it may also be jeopardized for future projects. Policy breaches that become widely known may erode the confidence of both funders and participants so that continued research or practice is impossible.

But the duties described in policies adopted by funders can also become legally enforceable, either when they are explicitly made part of a contract (e.g. between researcher and funding agency), or when courts draw on the standards they establish to determine whether a defendant has met the standard that should be expected of a reasonable researcher or practitioner to determine liability in tort law or delict.[2]

Although certain laws, notably the US Common Rule, may require that medical research projects submit a detailed research proposal to a research ethics body and obtain prior approval, the requirement is commonplace in the realm of policy. Because the objectives of funding policies are not limited to protecting participants—the sole aim of medical and data-privacy law—and because they also seek to foster a context in which medical research will thrive, they contain some unique requirements not found in law. The following section discusses one such topic, open-data requirements.

[2] In the common law and civil law traditions, respectively.

14.2 Open Data

Data-sharing duties have rapidly proliferated in the policies adopted by funding agencies that they impose on grant recipients, although these duties have not found their way into data-privacy laws. This section explains the rationale behind open data policies, describes some of the obligations that apply to research data, notes the presence of sharing repositories, and finally discusses the extension of this current into open publishing.

14.2.1 Rationale

A recent report by the Expert Advisory Group on Data Access (EAGDA), a joint initiative of four of the largest UK research institutions, lists four factors that favour making research data openly available (see Fig. 14.2).

These factors weigh particularly heavily in bioinformatics and the -omics fields. The data in question is rich and multidimensional to the point that it is difficult to imagine ever exhausting its research potential. Research methodologies like genome-wide association studies (GWAS) rely on increasingly large sample sizes: the larger the better (McCarthy et al. 2008). The cost of collecting—let alone sequencing—this data directly from large numbers of people anew for each and every research initiative would be prohibitive.

Despite the strong trend toward data sharing, open-data requirements cannot be absolute. The exception that proves the rule is the Personal Genome Project (PGP), which makes the genetic data of 3500 volunteers freely available for download on its website (personalgenomes.org). But genetic research projects generally cannot meet their objectives without guaranteeing privacy protection to participants. This may also be true for the PGP, whose aim is to sequence and publicize the complete genomes and medical records of 100,000 volunteers.

1. The scale of datasets being collected has grown dramatically, and these datasets are assembled at significant cost.
2. It will usually not be possible for one group to analyse these data exhaustively, and there will often be significant potential for the data to be used to answer questions distinct from the original research questions of the data producers.
3. Developments in information technologies are transforming the ease with which large datasets can be shared, linked and analysed.
4. Both those who volunteer their data and samples for research, and those who pay for that research, hope for progress towards useful and eventually applicable results for human health and other societal benefits to be as rapid as possible. Indeed, there is a clear ethical requirement for efficient use of data from human research participants.

Fig. 14.2 'Drivers for data sharing' listed by the UK Expert Advisory Group on Data Access (Expert Advisory Group on Data Access 2015)

Genetic researchers have their own concerns regarding open data, primarily the fear that before they have the chance to publish their findings, their collected data will be used by rival researchers who will publish their results first. The most common mechanism to address this concern has been to mandate embargo periods, during which researchers temporarily holds exclusive publication rights over 'their' data. The embargo mechanism, however, has proven difficult to enforce in practice, and thus appears less often in recent data-sharing policies.

Enforcement difficulties were explicitly cited by the US National Institutes of Health (NIH), for example, as the reason abandoning embargo periods were abandoned in the 2014 *Genomic Data Sharing Policy* (GDS), which now applies to all large-scale, NIH-funded genomics research (National Institutes of Health 2014a).

14.2.2 Extent of the Duty

Open data obligations are about more than simply making research data available. Funding agencies commonly require that applicants include a data-sharing plan with their research funding proposal. Policies may also require that the data meet standards for quality and interoperability, and almost always encourage or even require that researchers release their data as rapidly as possible. The accepted delay for release of research data can be as little as 24 hours after they are generated, following the recommendation of the 'Bermuda Principles' put forward in 1996 by leaders in the Human Genome Project (Marshall 2001). A minor trend in the reverse direction has emerged, for example in the 2014 GDS policy, in which the NIH sought to account for its elimination of the embargo period by pushing some of its data-release deadlines back to the time of initial publication (National Institutes of Health 2014a).

The GDS policy provides detailed guidance regarding the NIH's expected deadlines. A slightly simplified version of the table it provides appears in Fig. 14.3, which is divided both between data submission and data publication deadlines, as well as into five distinct levels the NIH distinguishes based on the amount of processing and analysis that have been carried out on the data.

14.2.3 Repositories for Data-Sharing

To make compliance with their mandatory data-sharing requirements easier, funding agencies sometimes provide researchers with technological resources to assist in the process, and in particular have established repositories to which the data can be submitted for future access for secondary research. The NIH database of Genotypes and Phenotypes (dbGaP), a repository for individual-level data, is likely the most well-known of these. Data-sharing policies sometimes require that the data be submitted to a repository that has been specifically approved by the funding agency. The GDS policy is one such example, although it additionally allows

	General Description	Example Data Types	Data Submission Expected	Data Release Timeline
Level 0	Raw generated data	Instrument image data	Not expected	
Level 1	Initial sequence reads	DNA sequencing reads, ChIP-Seq reads	By publication time for non-human, de novo data Not expected for human data	
Level 2	After initial analysis or computation to clean data and assess quality	DNA sequence alignments to a reference sequence	Within 3 months of data generation, for human data By publication time, for non-human data	Within 6 months of acceptance for publication or data submission, whichever occurs first, for human data By publication time, for non-human data
Level 3	Analysis to identify genetic variants, gene expression patterns, or other features	SNP or structural variant calls, expression peaks, epigenomic features	Within 3 months of data generation, for human data By publication time, for non-human data	Within 6 months of acceptance for publication or data submission, whichever occurs first, for human data By publication time, for non-human data
Level 4	Final analysis relating genomic data to phenotype or other biological states	Genotype-phenotype relationships, relationships of epigenomic patterns to biological state	As analyses are completed, for human data By publication time, for non-human data	Data released with publication, for human data No later than the time of initial publication, for non-human data

Fig. 14.3 Data submission and release deadlines (Adapted and abridged from a supplement to the *NIH Genomic Data Sharing Policy* for the five general levels of data it specifies (National Institutes of Health 2014b))

researchers to submit their data to external repositories, so long as these include privacy and security features that meet the policy's requirements (National Institutes of Health 2014c).

14.2.4 Open Publishing

Funding agency policies also frequently apply the 'open' ethos more broadly, and require not only that researchers' data be made available, but also the academic analysis they ultimately publish. This trend has been made possible the proliferation of open-access academic journals. In Canada, for example, the *Tri-Agency Open Access Policy on Publications* not only requires the submission of bioinformatics data to a public database in certain circumstances, it also mandates that *any* funding from the country's three principal scientific research agencies[3] comes with the obligation that the funding recipient will 'ensure that any peer-reviewed journal

[3] Namely, the Canadian Institutes of Health Research (CIHR); the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC).

publications arising from Agency-supported research are freely accessible within 12 months of publication' (Government of Canada 2015).

14.3 Consent

Informed consent has been a fundamental principle of healthcare and research for decades. Indeed, the topic dominated discussion in medical literature during the last half of the twentieth century (Manson and O'Neill 2007), and is seen as the essential mechanism for protecting patients and participants. Health care practitioners who provide personalized medicine to patients will be held to informed consent requirements, just as are other providers of health care.

But informed consent duties are usually less strict for secondary use of data or materials, for example, when research teams carry out studies on materials in biobanks or information in genomic data sharing repositories. This section focuses on consent to secondary research, and then also discusses dynamic consent, which has been proposed and has recently begun to be adopted in attempt to breathe new life into consent practices.

14.3.1 Secondary Use

Secondary research is difficult to reconcile with the normal approach to informed consent, which would require the initial participants to re-consent, which 'is costly and time-consuming, and difficulty in locating people can result in high drop-out rates' (Kaye et al. 2015). A variety of policy and legislative responses have emerged to decrease the intensity of specific consent requirements in the context of secondary use.

One approach is to simply abandon the consent requirement where it proves too onerous. Singapore's *Personal Data Protection Act 2012* (PDPA), for example, allows the use of personal data without consent for research when re-consent would be 'impracticable'. (To drop informed consent, the PDPA additionally requires that the research could not be accomplished without the data, it imposes limits on linkage with other data, and requires that the data subjects not be contacted.) In a guidance document, Singapore's Personal Data Protection Commission explained that the impracticability condition should be considered to be satisfied, for example, where the data was 'collected many years ago', because in that case the data subjects may have died or moved to another country in the intervening time (Personal Data Protection Commission (Singapore) 2014).

A second strategy is to eliminate the need for re-consent by seeking consent from participants at the moment when they initially consent to participate that is broad enough to also allow it to satisfy the consent needed for their participation in potential future research. Volunteers in the PGP, for example, whose data is available to any researcher or hobbyist who cares to make any conceivable use of

them, must consent to all potential future uses of their data before they are included in the open, online PGP repository. Somewhat similarly, the NIH's GDS policy 'expects investigators generating genomic data to seek consent from participants for future research uses and the broadest possible sharing' (National Institutes of Health 2014a).

But consent can easily become too broad. Blanket consent to any possible future research is inconsistent with many law and policy protections. Even when these protections allow departures from full informed consent, the tradeoff is usually an increase of other obligations, such as external monitoring and governance requirements.

One of the fundamental principles of the 1980 *OECD Privacy Guidelines*—which have influenced nearly every other data-privacy law that currently exists—requires that whenever personal data is collected, the purposes of collection are specified, and that any subsequent use of the data must be limited to those purposes. If the purpose of collection is stated too broadly, for example if the purpose is simply to allow participation in future research, this may prove to be insufficient to satisfy the specification requirement. Under the HIPAA Privacy Rule, for example, unless a research team member 'anticipated and adequately described the purposes of the secondary research in the initial authorization received from a patient, that initial authorization may not constitute authorization for the use of identifiable registry data for secondary research purposes' (United States Agency for Healthcare Research and Quality 2007).

Similarly, Article 6 of the current draft of the forthcoming EU Regulation prescribes a general prohibition on secondary use '[w]here the purpose of further processing is incompatible with the one for which the personal data have been collected', with only few exceptions, although this provision has been among those in the EU Regulation that have been most actively contested.

A third approach is to allow secondary use without fresh consent when privacy guarantees in place are likely to prevent harm to the participant that might flow from data use. Canada's *Tri-Council Policy Statement* (TCPS) adopts an approach similar to that of Singapore's PDPA, but additionally requires that appropriate privacy safeguards are in place. Rather than stating what, precisely, those privacy safeguards must be, the TCPS leaves the competent research ethics body the discretion to consider the question according to each particular set of circumstances. The US Common Rule implements the privacy guarantee approach in a more rigid manner. It simply exempts data that have been anonymized from having to conform to its requirements, by deeming research on anonymized data not to involve human subjects. In a somewhat similar way, the HIPAA Privacy Rule also allows secondary use of anonymized data.

14.3.2 Dynamic Consent

One additional strategy that might be leveraged to address the difficulties with secondary consent is the adoption of dynamic consent mechanisms, although no

prominent laws or policies currently explicitly require that they be used. The strategy is, however, being discussed with enthusiasm as a means to begin to address shortcomings associated with the traditional approach to informed consent more broadly (Erlich et al. 2014). The existing, standard notice-and-consent practice is characterized by lengthy consent forms that are presented to participants at the outset of their involvement research. The forms tend to leave the participant little meaningful choice, beyond the initial decision between whether to accept the conditions it describes, or to opt out of the research altogether. Critics liken this process to the lengthy terms and conditions often found in online contractual agreements, which invariably end with a single button marked 'I agree', which a person can choose to either click, or not.

If we are indeed entering into an era of *personalized medicine*, advocates of dynamic consent ask, why not also one of *personalized consent*? 'If biobank research is open-ended and ongoing then information technologies offer the possibility for participant involvement similarly to extend through time' (Kaye et al. 2015). The approach is most compelling where the participants' and patients' internet access is not overly hindered, either by technological, cultural, or educational barriers.

Research participants each have unique desires and expectations related to their research. Some may be comfortable with their data being shared for research into a specific disease, but feel that participation in unrelated research is not worth the privacy risks. Others may want their data to be available for a wide variety of medical research, but be opposed to their data being acquired or used by pharmaceutical corporations. Still others may want to prevent their personal health information from being used in studies that open it to a greater risk of government or law enforcement surveillance programs.

Dynamic consent strategies can allow not only for these decisions to be made by the participant and respected by researchers, especially in the clinical setting, but also allow for evolution over time of both the available options and preferences themselves. They can also allow participants' preferences to travel with their data samples. The approach seems to more fully embody and give meaning to the longstanding expectation that 'researchers will comply with any known preferences previously expressed by individuals about any use of their information' (TCPS).

Whether or not dynamic consent ultimately continues to expand in practice, consent will continue to retain its central role in medical practice despite undergoing significant changes in evolving contexts (Expert Advisory Group on Data Access 2015).

14.4 Anonymization

Until relatively recently, privacy experts invested a significant portion of their efforts into techniques to achieve data *anonymization* (or *de-identification*, which is sometimes used as a synonym, and other times, as a broader concept also encompassing pseudonymization). But a series of published re-identification

attacks has led to vigorous debate and reappraisal of the merits of anonymization, and health professionals and privacy experts have now increasingly been driven toward alternative strategies.

The basic practice of anonymization can be illustrated by considering aggregate statistics. Even if thousands of people's personal data must be mobilized to determine that Berlin has a population of 3½ million, that statistic itself reveals effectively nothing about any of the city's specific residents. Even if the statistic relies on a great amount of personal information, it is itself an anonymized datum.

Anonymization, however, is commonly carried out without aggregation. The paradigmatic example is an operation on set of records, each of which relates to a single person, which excizes or obfuscates enough information in each record to make it becomes impossible to use the resulting data set to identify any of the people initially connected to the data.

Data-privacy legislation usually addresses anonymization only implicitly: The laws usually restrict their scope so that they have no application to information in general, but only to *personal information*,[4] defined as information about an identifiable individual. Information that cannot identify an individual—such as the statistics mentioned above, or data that has otherwise been anonymized—falls outside of this scope and is therefore not subject in any way to data-privacy legal or policy protections. Some specific data-privacy protections also explicitly state that they do not apply to data that has been anonymized.

This section first describes different legal standards that data has to meet in order to be considered properly anonymized. It then explains why anonymization as a technique has fallen into disfavour among privacy exports and health professionals alike. The section finally discusses the legal implications of some of the new approaches that are beginning to occupy the place formerly held by traditional anonymization techniques.

14.4.1 Thresholds

Perfect anonymization is impossible: 'Data Cannot be Fully Anonymized and Remain Useful' (Dwork and Roth 2014). Perhaps unsurprisingly then, the legal and policy requirements for data to be considered properly anonymized vary. Different thresholds are sometimes deliberately specified depending on the use that will be made of the data, on its sensitivity, or on a combination of both factors. This follows the principle that the degree of anonymization should be proportionate to the intensity of potential harms that might result from misuse of the data, and the likelihood that those harms will, in fact, materialize.

Coding or *pseudonymization* is a strategy related to, but distinct from, anonymization. The practice allows data sets to retain an identifier whose purpose is to allow the re-identification of data which have otherwise been anonymized, but

[4] Or any of various synonyms used, such as 'identifying data'.

to allow this only by people with access to a separate, private data set that links the identifiers back to individually identifying information.

While the HIPAA Privacy Rule's conception of 'de-identified' data, as discussed below, encompasses coded data, the EU Data Protection Directive excludes it by defining personal data as data relating to a person 'who can be identified, directly or indirectly, in particular by reference to an identification number'.

The various legal and policy definitions of personal data almost invariably remain at a highly abstracted level. In the broadest terms, personal data can be cast either *narrowly*, as occurs in definitions that include only data in which a person's identity is 'readily ascertainable', or *broadly*, as occurs in definitions that include any data for which it is reasonably foreseeable that the data (either alone or in combination with other data sets) will allow an individual to be identified.

One notable exception to the trend of defining anonymization in general terms is the HIPAA Privacy Rule, whose definition delves into an unusual level of technical detail. The Privacy Rule provides two alternative procedures, either of which allows data to be considered de-identified for HIPAA's purposes. The first option is to obtain a detailed written opinion from a statistician assuring that the re-identification risk that can reasonably be anticipated is 'very small'. The second option, sometimes referred to as the HIPAA 'Safe Harbor', requires that each of seventeen specified fields be removed from every record in the data set (see Fig. 14.4).

(A) Names;

(B) All geographic subdivisions smaller than a State ... except for the initial three digits of a zip code if ...

(1) The geographic unit formed ... contains more than 20,000 people; and

(2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

(C) All elements of dates (except year) ... directly related to an individual ... ; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses;

(G) Social security numbers;

(H) Medical record numbers;

(I) Health plan beneficiary numbers;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) Device identifiers and serial numbers;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images;

Fig. 14.4 The seventeen HIPAA privacy rule de-identification fields

(i) Names;
(ii) Postal address information, other than town or city, State, and zip code;
(iii) Telephone numbers;
(iv) Fax numbers;
(v) Electronic mail addresses;
(vi) Social security numbers;
(vii) Medical record numbers;
(viii) Health plan beneficiary numbers;
(ix) Account numbers;
(x) Certificate/license numbers;
(xi) Vehicle identifiers and serial numbers, including license plate numbers;
(xii) Device identifiers and serial numbers;
(xiii) Web Universal Resource Locators (URLs);
(xiv) Internet Protocol (IP) address numbers;
(xv) Biometric identifiers, including finger and voice prints; and
(xvi) Full face photographic images and any comparable images.

Fig. 14.5 Protected health information that excludes these direct identifiers of the person to whom the data relates and of their relatives, employers, or household members qualifies as a Limited Dataset under the HIPAA Privacy Rule

HIPAA Safe Harbor anonymization additionally requires the removal of any other unique identifying number, with the exception of an optional re-identification number (which would thus result in a coded, rather than an anonymized, data set). The Safe Harbor's final requirement is that the person carrying out the anonymization must not have 'actual knowledge that the information could be used alone or in combination with other information to identify an individual'.

A less strict variation on Safe Harbor anonymization, called a limited data set, is also described by HIPAA. The use of limited data sets, as a tradeoff, requires that researchers sign a data-use agreement subjecting them to additional restrictions on how the data may be used and to whom it may be disclosed. The limited data set anonymization fields appear in Fig. 14.5. Limited data sets are most commonly used by researchers who want to analyze the additional data fields that can legitimately be retained, such as dates and five-digit zip codes.

14.4.2 Anonymization's Fall into Disfavour

The Safe Harbor's straightforward anonymization instructions may seem appealing when compared with a duty to anonymize data according to the vague standard requiring that it is no longer reasonably foreseeable that they will one day allow

re-identification. But the apparent simplicity of the Safe Harbor is a false promise, especially where genetic data is concerned, which is notably absent from the anonymization fields listed in Figs. 14.4 and 14.5.

Because all but the shortest genetic sequences are high dimensional, these data are generally thought to be impractical to anonymize (El Emam and Arbuckle 2013). Some have argued that to comply with the Safe Harbor rules genetic information must be removed because it itself constitutes a unique identifying number under the eighteenth HIPAA Safe Harbor identifier. Guidance issued in the intervening years has, however, failed to address the issue (Office for Civil Rights 2012). The Safe Harbor's 'actual knowledge' requirement should also generally prevent genetic data from being included in a data set, though that test is drafted to depend on the mindset of the person doing the anonymization rather than on reasonable expectations of re-identifiability. The requirement is thus often ignored in practice (El Emam and Arbuckle 2013).

Beyond the challenges posed by genetic data, the HIPAA Safe Harbor is an illustration of the broader problems with attempts to set out a detailed anonymization procedures in law that do not take into account specific contexts. A 2009 report by the US Institute of Medicine found a number of failings with the HIPAA Privacy Rule, and emphasized that HIPAA's rigid procedure is simultaneously too strict and not strict enough. It is indeed trivial to construct an example data set for which the Safe Harbor both allows re-identification and also requires that data be unnecessarily removed.

Not only is HIPAA's approach to anonymization less than optimal, but the broader practice of anonymization itself has now increasingly fallen out of favour as an effective means of privacy protection, particularly when it comes to high dimensional data such as genomic sequences. Existing techniques are able to re-identify an individual given as few as thirty independent single nucleotide polymorphisms (SNPs) (El Emam and Arbuckle 2013), and so to anonymize any genetic sequence with confidence would often require obliterating most of the data, along with its research value. In the same vein, anonymization is coming to be seen as unhelpful to translational medicine, which relies on linkage between different data sets, and is impossible once anonymization effectively sterilizes them.

If debate about the continued relevance of anonymization has not been completely settled, perhaps it is because its remaining defenders have already conceded so much. It is increasingly rare for anonymization to be used as a privacy safeguard in practice on its own, without being supported by other mechanisms. After researchers showed that data in dbGaP could be re-identified despite having been anonymized according to HIPAA, the NIH converted dbGaP into a controlled- rather than open-access repository (Homer et al. 2008; National Institutes of Health 2008). In the UK, EAGDA now similarly recommends alternative protections such as access controls (Expert Advisory Group on Data Access 2013), although new re-identification attacks continue to be described in the literature (Cai et al. 2015). In 2014, the New Zealand Privacy Commissioner suggested addressing anonymization's weaknesses by going so far as to make it illegal to attempt to re-identify data (Edwards 2014).

14.4.3 Successors

Data-privacy experts are now turning away from anonymization and have begun to explore emerging alternative approaches to privacy protection, which sometimes include the potential to achieve provable security. The goal is no longer to anonymize datasets as much as possible so that they can be shared as widely as possible. Instead, many of the new strategies are based on cryptographic methods which aim to allow genomic research studies to be carried out without the need for any of the raw research data itself to ever need to be disclosed. At the forefront of these techniques are homomorphic encryption, secure multiparty computation, and differential privacy.

Homomorphic encryption is an attractive approach in cases where a third party is made responsible for storage and computation whose access to the data would itself be a privacy risk, such as in the context of the increasingly prevalent practice of genomic research using cloud computing services (Lauter et al. 2014). Genomic data is uploaded to the cloud in an encrypted form, and homomorphic encryption then allows researchers to submit calculations to have the cloud perform on the encrypted data and to ultimately receive the encrypted result, all the while maintaining the data in its encrypted form so that it remains unreadable to other parties, including the cloud service provider itself.

Secure multiparty computation is a related strategy. In this case the data set is split between multiple parties so that each one holds only a fraction of the overall data to be analyzed. Cryptographic methods then allow researchers the parties to collectively carry out calculations on the full data set without any individual party having to reveal any of their own raw data (Kamm et al. 2013). Similarly, techniques such as DataSHIELD allow researchers to perform aggregate calculations and studies on data sets held by third parties without the need to reveal any raw data to the researcher (Wolfson et al. 2010).

Differential privacy offers perhaps the most promise of all of these new methods, and is used in contexts of statistical aggregation. This method aims to mathematically determine whether an individual's decision to participate in a given study will have any effect on their privacy. Dwork and Roth describe differential privacy as a 'promise' that those holding data make to a data subject: 'You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available' (Dwork and Roth 2014).

These techniques have not yet made their way into laws and policies, and even though they are largely departures from anonymization, the legal analysis of their use must consider the rubric of personal information. This is so because despite the mathematical proofs that have been published demonstrating some of the methods' abilities to securely protect privacy, none has yet resulted in a generalizable method to ensure privacy protection in practice. Current practical methods of homomorphic encryption, for example, still require an 'assum[ption] that all [collaborating] entities behave *semi-honestly*' (Lu et al. 2015). Because calculations in secure

multiparty computation are always based on the raw data, the results they produce must reveal a degree of private information. For reasons such as these, it will remain necessary to ask whether it is reasonably foreseeable that the information revealed by these techniques allows individuals to be identified are likely to remain relevant. If the answer is yes, the associated data-privacy law restrictions will continue to apply.

14.5 Cross-Border Transfer

Concerns about cross-border transfer of data have grown considerably following Edward Snowden's revelations about the existence of widespread electronic surveillance programs. Both legal and policy restrictions now exist on cross-border transfer of personal information have increased in intensity and expanded in number. The rapidly expanding use of cloud computing in bioinformatics fields has also added to these concerns, given what is seen as the inherent borderlessness of cloud technologies.

The overarching concern with cross-border transfers and outsourcing of personal data is that these can place the data in contexts where they may be exposed to more serious privacy risks, and in particular, risks that the data holder is required not to expose them to. Because Canada's *Tri-Council Policy Statement*, for example, requires that researchers 'avoid being put in a position of becoming informants for authorities' (Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada 2014), this requires that researchers seek to avoid cross-border transfer to a jurisdiction known to engage in such surveillance programs.

In the broadest terms, most laws and policies aim to allow cross-border transfer and outsourcing when this will not significantly undermine data privacy. Two general approaches have emerged in data-privacy law with the objective of achieving this aim in the context of cross-border personal data transfer.

The *accountability* approach requires the entity transferring data to ensure that it will enjoy a similar or greater degree of protection in the hands of the specific entity to which the data is transferred in another jurisdiction. Canada has adopted this approach in its *Personal Information Protection and Electronic Documents Act* (Canada 2000).

The *adequacy* approach, in contrast, allows cross-border transfer only if the target jurisdiction has previously been deemed adequate by the data-privacy authority tasked with making such determinations. Adequacy determinations are required for cross-border transfer by the EU *Directive*, and the same approach will be retained in the new EU Regulation, and existing adequacy determinations will remain in force. Data-privacy laws in thirteen different jurisdictions have currently been approved by the European Commission as providing adequate protection.

Other laws, particularly since the Snowden revelations, have imposed blanket prohibitions on transfer or storage outside the jurisdiction, which may be subject to exceptions. The law of the Canadian province of British Columbia, for example, has long included such a blanket provision, which prohibits public bodies from storing personal information outside of Canada (British Columbia 1996). In 2014, however, the Information & Privacy Commissioner of that province published updated guidance stating that it is possible for public bodies to store personal data outside of the country without violating the law if the data are protected by a data security technique called tokenization (Office of the Information & Privacy Commissioner for British Columbia 2014). Tokenization is somewhat similar to coding. In this case, it would allow data sets to be stored outside of Canada so long as any personal information has been replaced by a 'token'. The token allows the personal information it represents to then be retrieved using a separate data set known as a 'crosswalk table', which must be stored in Canada.

14.6 Genetic Nondiscrimination

Laws and guidelines have recently proliferated that prohibit certain forms of discrimination on the basis of genetic information. The US *Genetic Information Nondiscrimination Act of 2008*, for example, was discussed in the Introduction to this chapter, especially with respect to gaps in the protection it provides against certain 'milder' forms of genetic discrimination.

But like many other genetic nondiscrimination laws, GINA is also limited in terms of who it prevents from engaging in discrimination. GINA applies only to insurance and employment sectors. But this does not mean that researchers in clinical bioinformatics are free to disregard these laws, which often contribute to determining the risks of discrimination to which research participants are exposed. The Philippine *National Ethics Guidelines for Health Research 2011*, for example, explicitly require that research projects involving genetic data contend with the issue (Philippine National Health Research System 2011):

> There is potential harm to research participants arising from the use of genetic information, including stigmatization or discrimination. Researchers should take special care to protect the privacy and confidentiality of this information.

Beyond providing these privacy and confidentiality protections, clinical bioinformaticians must be aware of any participant or patient interaction that could reasonably increase the risk of becoming subject to genetic discrimination. For example, although Australia's *Insurance Contracts Act 1984* prohibits insurance companies from requiring that a customer undergo genetic testing, if the customer has already had a genetic test, and even if they simply know the results a family member's test, the results must be declared before entering into a new insurance contract (Liddell 2002).

Article 11 – Non-discrimination

Any form of discrimination against a person on grounds of his or her genetic heritage is prohibited.

Article 12 – Predictive genetic tests

Tests which are predictive of genetic diseases or which serve either to identify the subject as a carrier of a gene responsible for a disease or to detect a genetic predisposition or susceptibility to a disease may be performed only for health purposes or for scientific research linked to health purposes, and subject to appropriate genetic counselling.

Article 13 – Interventions on the human genome

An intervention seeking to modify the human genome may only be undertaken for preventive, diagnostic or therapeutic purposes and only if its aim is not to introduce any modification in the genome of any descendants.

Article 14 – Non-selection of sex

The use of techniques of medically assisted procreation shall not be allowed for the purpose of choosing a future child's sex, except where serious hereditary sex-related disease is to be avoided.

Fig. 14.6 Chapter IV of the Council of Europe's 1997 *Oviedo Convention*, which sets out basic protections with respect to the human genome (Council of Europe 1997)

In some jurisdictions, clinical bioinformaticians themselves are additionally directly subject to prohibitions on genetic discrimination. One of the earliest genetic nondiscrimination laws, for example, the Council of Europe's *Oviedo Convention*, does not limit its scope to any particular categories of potential discriminators (Fig. 14.6).

14.7 Conclusion

Although legal and policy duties regulate clinical bioinformaticians in areas beyond those discussed in here—including transaction logging, data-privacy breach notification, risk assessment, and reporting of incidental findings, among others—this chapter provided an introduction to five areas of key importance. Some were chosen because they hold a fundamentally important place in the legal and policy frameworks, while others have been the subject of extensive expert debate and discussion. In either case, familiarity with these concepts is helpful in contending with the broader issues. The discussion focused on law and policy from several world regions, to illustrate the ways in which positions on each question can either converge or deviate around the globe.

The fundamental tension in the field, which presents itself at every turn, remains finding the optimal balance between privacy protection and the facilitation of medical research and care. What often appear to be new areas of debate—such as the question of open data, or even of the continued relevance of anonymization techniques—each soon reveal themselves to be manifestations of that same initial underlying tension. If the issue of open data in genomics were entirely independent of this tension, the most vocal advocates of open bioinformatics research data might

be expected to be seen applying the idea of a genomic commons to the issue of genetic patents, arguing to limit these or eliminate them altogether, but efforts in this direction are, if anything, currently declining (Contreras 2015). Thankfully, robust promotion of both health research and of privacy protections do not always have to be played off against one another in a zero-sum game. Many of the techniques described in this chapter that have only recently begun to be explored and that have yet to be internalized by law and policy at all, such as homomorphic encryption and dynamic consent, appear to have the potential to promote both.

Disclosure Statement Funding for this work was provided by the Canada Research Chair in Law and Medicine.

References

British Columbia. Freedom of information and protection of privacy act. Revised Statutes of British Columbia, chapter 165. Queen's Printer BC; 1996.

Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, Zhou S. Deterministic identification of specific individuals from GWAS results. Bioinformatics. 2015;31(11):1701.

Canada. Personal information protection and electronic documents act. Statutes of Canada, chapter 5. Queen's Printer for Canada; 2000.

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada. Tri-council policy statement: ethical conduct for research involving humans. Ottawa: Secretariat on Responsible Conduct of Research; 2014.

Contreras JL. NIH's genomic data sharing policy: timing and tradeoffs. Trends Genet. 2015;31(2):55.

Council of Europe. Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: convention on human rights and biomedicine. Strasbourg: Council of Europe; 1997.

Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comp Sci. 2014;9(3–4):211.

Edwards J. Privacy and big data: presentation by privacy commissioner John Edwards. New Zealand Privacy Commissioner; 2014.

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

Erlich Y, Williams JB, Glazer D, Yocum K, Farahany N, Olson M, et al. Redefining genomic privacy: trust and empowerment. PLoS Biol. 2014;12(11):e1001983. doi:10.1371/journal.pbio.1001983.

Expert Advisory Group on Data Access. Governance of data access. 2015. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_peda/documents/web_document/wtp059343.pdf

Expert Advisory Group on Data Access. Statement for EAGDA funders on re-identification. 2013. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp055972.pdf

Government of Canada. Tri-agency open access policy on publications. 2015. http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4(8):1000167.

Horvitz E, Mulligan D. Data, privacy, and the greater good. Science. 2015;349(6245):253.

Hunter v. Southam. 11 Dominion Law Reports, 4th Series, 641. 1984.
Kamm L, Bogdanov D, Laur S, Vilo J. A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics. 2013;29(7):886.
Katz v. United States. 389 US 347. 1967.
Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eu J Hum Gen. 2015;23:141.
Lauter K, López-Alt A, Naehrig M. Private computation on encrypted genomic data. 2014. http://research.microsoft.com/pubs/219979/genomics.pdf
Liddell K. Just genetic discrimination? The ethics of Australian law reform proposals. Univ NSW Law J. 2002;25(1):160.
Lu W, Yamada Y, Sakuma J. Efficient secure outsourcing of genome-wide association studies. IEEE CS Security and Privacy Workshops. 2015.
Manson NC, O'Neill O. Rethinking informed consent in bioethics. Cambridge: Cambridge University Press; 2007.
Marshall E. Bermuda rules: community spirit, with teeth. Science. 2001;291(5507):1192.
McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356.
National Institutes of Health. Modifications to genome-wide association studies (GWAS) data access. 2008. http://gds.nih.gov/pdf/Data%20Sharing%20Policy%20Modifications.pdf
National Institutes of Health. NIH genomic data sharing policy. 2014c. http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf
National Institutes of Health. Notice number: NOT-OD-14-124. 2014a, 27 August. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html
National Institutes of Health. Supplemental information to the National Institutes of Health genomic data sharing policy. 2014b. http://gds.nih.gov/PDF/Supplemental_Info_GDS_Policy.pdf
Office for Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html
Office of the Information & Privacy Commissioner for British Columbia. Updated guidance on the storage of information outside of Canada by public bodies. 2014, 16 June. https://www.oipc.bc.ca/public-comments/1649
Personal Data Protection Commission (Singapore). Advisory guidelines for the healthcare sector. 2014, 11 September.
Philippine National Health Research System. National ethical guidelines for health research. 2011. http://www.ethics.healthresearch.ph/index.php/phoca-downloads/category/4-neg?download=9:pub-ethics-guidelines-2011
Rothstein MA. Currents in contemporary ethics: GINA, the ADA, and genetic discrimination in employment. J Law Med Ethics. 2008;36(4):837.
Tobin SL, Lee SSJ, Greely HT, Ormond KE, Cho MK. Not a loophole: commercial exploitation of an IRB error. Comment on: Gibson G, Copenhaver GP. Consent and internet-enabled human genomics. PLoS Genet. 2010;6(6):e1000965.
United States Agency for Healthcare Research and Quality. Registries for evaluating patient outcomes: a user's guide. Rockville: US Department of Health and Human Services; 2007.
Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol. 2010;39(5):1372.

Mark Phillips holds an LL.B. and a B.C.L. from McGill University's Faculty of Law, as well as a B.Sc. (Honours) in Computer Science from the University of Manitoba. His work at the Centre of Genomics and Policy is focused on comparative analyses of data protection, privacy, and cloud computing laws and policies, particularly as they relate to bioinformatics. He is a former editor of both the *McGill Journal of Law and Health* and of the *McGill Law Journal* and has published several peer-reviewed articles and book chapters. His research interests also include computer-assisted legal research methodologies, mental health and disability, and law and social movements.