See ARTICLES pages 529, 546, and 554

Coding of DNA Samples and Data in the Pharmaceutical Industry: Current Practices and Future Directions—Perspective of the I-PWG

MA Franc[1], N Cohen[1], AW Warner[2], PM Shaw[3], P Groenen[4] and A Snapir[5]; on behalf of the Industry Pharmacogenomics Working Group

DNA samples collected in clinical trials and stored for future research are valuable to pharmaceutical drug development. Given the perceived higher risk associated with genetic research, industry has implemented complex coding methods for DNA. Following years of experience with these methods and with addressing questions from institutional review boards (IRBs), ethics committees (ECs) and health authorities, the industry has started reexamining the extent of the added value offered by these methods. With the goal of harmonization, the Industry Pharmacogenomics Working Group (I-PWG) conducted a survey to gain an understanding of company practices for DNA coding and to solicit opinions on their effectiveness at protecting privacy. The results of the survey and the limitations of the coding methods are described. The I-PWG recommends dialogue with key stakeholders regarding coding practices such that equal standards are applied to DNA and non-DNA samples. The I-PWG believes that industry standards for privacy protection should provide adequate safeguards for DNA and non-DNA samples/data and suggests a need for more universal standards for samples stored for future research.

Pharmaceutical company–sponsored clinical trials (controlled phase III trials in particular) provide a unique opportunity to conduct productive pharmacogenomic research. To this end, sponsors have introduced processes to collect DNA samples to enable pharmacogenomic studies that address clinical, scientific, and regulatory issues in drug development. There are several documented instances of issues emerging during the course of clinical development that triggered important scientific and regulatory questions critical to understanding the benefit–risk profile of a compound even after study closure.[1] Provisions for storage of DNA samples for future research are therefore a critical aspect of pharmacogenomics in drug development. The challenge for study participants and institutional review boards (IRBs)/ethics committees (ECs) is to decide whether privacy protection measures are adequate and to ensure that the research is justified and the risk to the individual is minimized.

There has been some uncertainty about the risks of genetic research. This has stemmed, in part, from certain unique properties of DNA, a paucity of laws preventing discriminatory use of genetic data, fears regarding inappropriate disclosure of data, potential use of the data for commercial benefit, and uncertainty as to the interpretability and utility of information in the future as technology and knowledge evolve.[2] The protection of patients' privacy and rights is a fundamental requirement for the conduct of clinical trials.[3] Pharmaceutical companies are extensively regulated with respect to maintenance of patients' privacy and are responsible for ensuring that privacy is appropriately protected. The industry has a history of maintaining strong levels of protection of data through various methods, including the use of secure facilities and databases with restricted access, appropriate policies, standard operating procedures (SOPs), and training, to ensure compliance with privacy protections per good clinical practice[3] and other local regulations.

To alleviate concerns and allow pharmacogenomic research to move forward, the pharmaceutical industry has adopted various coding methods for samples and data. These methods have largely served to satisfy IRBs/ECs that may otherwise not allow DNA collection and banking and to conform to local laws and regulations that vary widely and are subject to interpretation by national authorities and IRBs/ECs.[4,5] The more complex coding

[1]Department of Pharmacogenomics, Johnson & Johnson Pharmaceutical Research and Development, Raritan, New Jersey USA; [2]Clinical Pharmacogenomics and Clinical Specimen Management, Merck Research Laboratories, Whitehouse Station, New Jersey, USA; [3]Department of Genetics, Merck Research Laboratories, Whitehouse Station, New Jersey, USA; [4]Molecular Design and Informatics/Exploratory and Translational Sciences, Merck Research Laboratories, Whitehouse Station, New Jersey, USA; [5]Translational Sciences, Orion Pharma, R&D, Turku, Finland. Correspondence: MA Franc (mfranc@its.jnj.com)

Received 26 July 2010; accepted 9 November 2010; advance online publication 23 February 2011. doi:10.1038/clpt.2010.306

methods have been predominantly applied to DNA samples stored for future research. Although these steps are well intentioned, whether they enhance privacy protection in the context of the pharmaceutical industry is being questioned.

Inconsistent and discrepant use of terminology for coding techniques has caused misunderstanding about how pharmaceutical companies (sponsors) protect patient privacy. In order to introduce consistency, the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Pharmaceutical Use (ICH) released the E15 guideline defining key terms for coding techniques.[6] The guideline, which leveraged previous Industry Pharmacogenomics Working Group (I-PWG)[7] and European Federation of Pharmaceutical Industries and Associations perspectives, provides definitions for four categories of pharmacogenomic sample/data coding: identified, coded (single and double), anonymized, and anonymous (see **Table 1**/**Figure 3**) and summarizes their implications (**Table 2**). Anonymous samples are not applicable to clinical trials and are not further discussed.

The key questions to be addressed are:

- Do additional coding procedures offer significant added privacy protection beyond standard coding procedures applied in pharmaceutical clinical trials, and can these perhaps impart a false sense of security?
- Do more restrictive coding procedures limit the potential utility of specimens, and are they necessary?
- Should DNA samples be coded any differently than other types of samples that are stored for future research?
- Are industry standards for protection of clinical data adequate and robust for storage of samples for future research?

Table 1 Summary of definitions of genomic data and sample coding categories adapted from ref. 6

Coding category	Summarized definitions
Identified	Labeled with personal identifiers such as name or identification numbers (e.g., social security or national insurance number)
Coded	Labeled with at least one specific code and do not carry any personal identifier numbers (e.g., social security or national insurance number)
Single-coded	Usually labeled with a single specific code and do not carry any personal identifiers. It is possible to trace the data or samples back to a given individual with the use of a single coding key
Double-coded[a]	Initially labeled with a single specific code and do not carry any personal identifiers. The data and samples are then relabeled with a second code which is linked to the first code through a second coding key. It is possible to trace the data or samples back to the individual by the use of both coding keys
Anonymized	Initially single- or double-coded but with the link between the subjects' identifiers and the unique code(s) being subsequently deleted. Once the link has been deleted, it is no longer possible to trace the data and samples back to individual subjects through the coding key(s). Anonymization is intended to prevent subject re-identification
Anonymous	Never labeled with personal identifiers when originally collected, and no coding key ever generated. Therefore, there is no potential to trace back genomic data and samples to individual subjects

[a]This method is also commonly referred to as "de-identified."

This paper attempts to address these questions by reporting current industry practices and opinions on DNA coding methods based on the results of a survey conducted by the I-PWG.

RESULTS

Survey respondents

Twenty-one companies contributed to the results of this survey (see Methods), which are summarized by Franc *et al.* in this issue.[1] Majority opinions and practices are highlighted below; the range of responses can be determined from tallied data in the **Supplementary Data** online.

DNA coding practices in the industry

Perceived higher risk of loss of privacy associated with DNA sampling. The majority of companies surveyed (68%) reported applying a more stringent level of coding to DNA samples/data as compared with other types of samples. Only one company reported also applying more stringent coding to non-DNA samples.

Reported coding practices. Companies were asked to refer to the ICH definitions when responding to the survey. Because the term "de-identified" is used by some pharmaceutical companies to refer to double-coding,[7] these terms were used in tandem. The most common coding practice for DNA storage was double coding/de-identification (65%). Only 10% of companies anonymized the data, and 20% single-coded them (**Figure 1**).

Personal identifiers. A key aspect of sample/data coding methods is protection of "personal identifiers." Whether a data element is considered to be a "personal identifier" currently varies across countries. Under the HIPAA (Health Insurance Portability and Accountability Act) privacy rule in the United States,[8] certain data elements have been enumerated as identifiers, and most of these are also considered identifiers under the 95/46/EC Directive in the European Union[9] (**Table 3**). The survey revealed differences in company definitions of personal identifiers, as shown in **Figure 2**. The majority of the companies said they use the identifiers set forth in the HIPAA privacy rule as an initial guide.

Securing and use of the secondary key. Of the companies that double-code/de-identify DNA data, the majority reported retaining the secondary key internally either by a party independent of the one responsible for double-coding/de-identification (38%) or by the same party but with restricted access (38%). A minority of the respondents (15%) relied on an external third party, and one relied on an internal third party not involved in the data analysis.

The most common reasons to use the secondary key to re-link to the original subject ID were to return data to subjects/

Table 2 Summary of implications of genomic data and sample coding categories adapted from ref. 6

Sample/data coding category	Link between subject's personal identifiers and genomic biomarker data	Traceability back to the subject[a]	Extent of subject's confidentiality and privacy protection
Identified	Yes (direct). Allows for subjects to be identified	Yes	Similar to general health-care confidentiality and privacy
Coded			
Single	Yes (indirect). Allows for subjects to be identified (by means of a single, specific coding key)	Yes	Standard for clinical research
Double	Yes (very indirect). Allows for subjects to be identified (by means of two specific coding keys)	Yes	Added privacy and confidentiality protection as compared to single coding
Anonymized	No. Does not allow for subjects to be re-identified, because coding key(s) have been deleted	No	Genomic data and samples no longer linked to subject, because coding key(s) have been deleted
Anonymous	No. Identifiers never collected and coding keys never applied. Does not allow for subjects to be identified	No	Genomic data and samples never linked to subject

[a]Traceability back to the subject affects ability to perform actions such as: sample withdrawal or return of individual genomic results at a subject's request; ability to perform clinical monitoring, subject follow-up, and addition of new data.

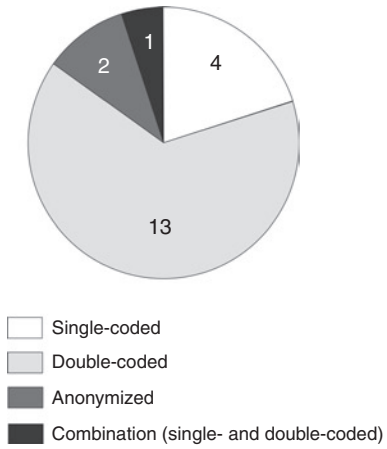

Figure 1 DNA coding practices used by survey respondents ($n = 20$). The numbers are number of responses.

physicians (44%) and for regulatory submission purposes (44%). Other reasons cited were to add data to the data set (38%), verify accuracy of the double-coding/de-identification process (13%), re-contact subjects for further study (6%), and destroy samples upon consent withdrawal (6%).

Link reconstruction. For double-coded/de-identified and anonymized data, the link between the original and new subject identifiers can be readily reconstructed by matching clinical data parameters across data sets (see Discussion; **Figure 3b**). Among companies that double-coded/de-identified, the most common approach to guard against this was to restrict access to both data sets (e.g., database password protection/IT systems). Another common approach was to prohibit link reconstruction in an SOP document. Some used a combination of approaches. Five companies used a statistical disclosure limitations methodology in which values are stored as ranges.

Practical implementation of DNA collection at a global level. Many companies (40%) had sometimes encountered requests from IRBs/ECs for procedures that differed from their standard procedures, 20% had encountered these often, 25% only rarely, and 10% never. An additional 5% said they did not know. Companies reported engaging in negotiation with IRBs/ECs with varying success depending on the nature of the request and the extent to which their processes could accommodate it. Twenty percent of the respondents reported having to often forego DNA collection. The feedback indicated that the requirements were generally related to coding methods and duration of storage, and that case-by-case concessions were made in order to obtain as many samples as possible.

Properties of coding methods. The respondents' opinions on the intrinsic properties of coding methods as they relate to privacy protection, study approval, subject participation, and submissibility of data are summarized in **Figure 4**. Opinions were divided regarding the ability of any one method to inherently ensure patient privacy; however, most of the respondents agreed that coding methods must be combined with other measures such as SOPs and training in order to be effective. Free-text comments revealed a strong sentiment that anonymization and double-coding/de-identification were difficult to implement, cumbersome, costly, and inefficient and offer little added privacy protection in the context of research in the pharmaceutical industry. One company explicitly stated that it did not apply "genetic exceptionalism" and that all samples, including DNA, were single-coded. Overestimation of the added level of protection afforded by double-coding/de-identification and anonymization methods was raised as a concern.

Recommendations of the survey respondents. When asked for recommendations for the minimal level of coding required for DNA sample/data storage in the current global environment, the majority of the companies (70%) felt that single-coding was sufficient for short-term storage (up to 1 year after the trial); for long-term storage, 60% recommended double-coding/de-identification (**Figure 5**).

When asked for recommendations for the minimal level of coding required for DNA sample/data storage moving forward, the majority (80%) agreed that single-coding was sufficient for

Table 3 Personal identifiers according to the Health Insurance Portability and Accountability Act (USA),[8] 95/46/EC Directive (European Union (EU)),[9] and ICH E15 (ref. 6)

Personal identifiers		
USA[8]	1	Names
	2	All geographic subdivisions smaller than a state
	3	All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; also all ages > 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"
	4	Telephone numbers
	5	Fax numbers
	6	Electronic mail addresses
	7	Social security numbers
	8	Medical record numbers
	9	Health plan beneficiary numbers
	10	Account numbers
	11	Certificate/license numbers
	12	Vehicle identifiers and serial numbers, including license plate numbers
	13	Device identifiers and serial numbers
	14	Web Universal Resource Locators (URLs)
	15	Internet protocol address numbers
	16	Biometric identifiers, including finger- and voiceprints
	17	Full face photographic images and any comparable images
	18	Any other unique identifying number, characteristic, or code
EU[9]		"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
ICH E15[6]		"...personal identifiers, such as name or identification numbers (e.g., social security or national insurance number)"

short-term storage, whereas opinions were equally divided about long-term storage, with 50% recommending single-coding and 50% recommending double-coding/de-identification.

DISCUSSION

Greater stringency regarding privacy protection practices related to DNA

Several factors have contributed to the adoption of stricter rules and practices for collection of specimens for pharmacogenomic research. These are based on ethical, logistical, and social issues and varying interpretations of the risk and value that the research might provide. The issues include (i) the scarcity of laws pertaining to discrimination on the basis of genetic profile, (ii) the altruistic nature of sample provision that offers no immediate direct benefit to the subject, (iii) certain unique properties of DNA, (iv) the lack of harmonization or best practices, (v) ideological arguments that question the legitimacy of seeking consent for future research when the exact details of such research are not known, (vi) lack of understanding of the value proposition of pharmacogenomics, and (vii) a misunderstanding of data-driven genomics approaches (e.g., genome-wide association studies).

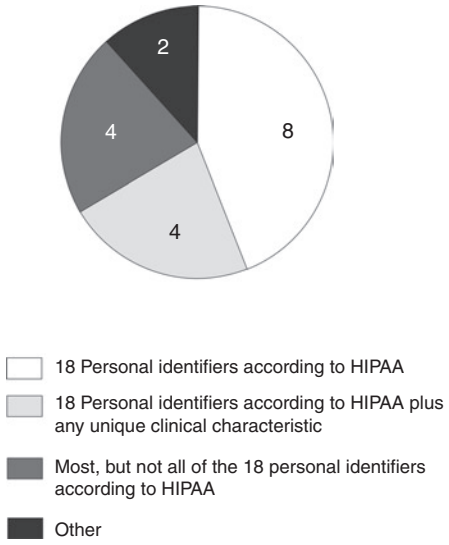

Figure 2 Various company definitions of the term "personal identifiers" as it relates to sample and data coding practices. Health Insurance Portability and Accountability Act (HIPAA) of 1996 (45CFR 164.514(b)(2)) ($n = 18$). The numbers are number of responses.

Unique properties of DNA

Every individual (except a monozygotic sibling) possesses a unique genome, which is largely unchanging over time. Although genetic data, in isolation, are not considered to be personally identifying,[10,11] it is conceivable that genetic data could be used to identify a person if combined with another data set that contains sufficient overlapping genetic data as well as personal identifiers.[11–13] DNA is also highly stable and can be maintained almost indefinitely; therefore, the information contained within the sample can be available for long periods of time. DNA is heritable, allowing predictions or inferences about blood relatives and relatedness. Finally, DNA information can be used for providing relatively accurate diagnoses/prognoses in highly penetrant Mendelian disorders. This has led to the inaccurate perception that all genetic information *per se* will provide highly accurate predictions about an individual's risk for all diseases, or other traits.

Although these properties of DNA are associated with certain risks to individual privacy when samples are stored or used, these risks are not effectively mitigated by additional sophisticated coding methods for research conducted by industry. Rather, other measures currently applied by the industry may be more effective at addressing these risks, including predefining the duration of storage, removing personal identifiers from samples/data used in research collaborations, defining the scope

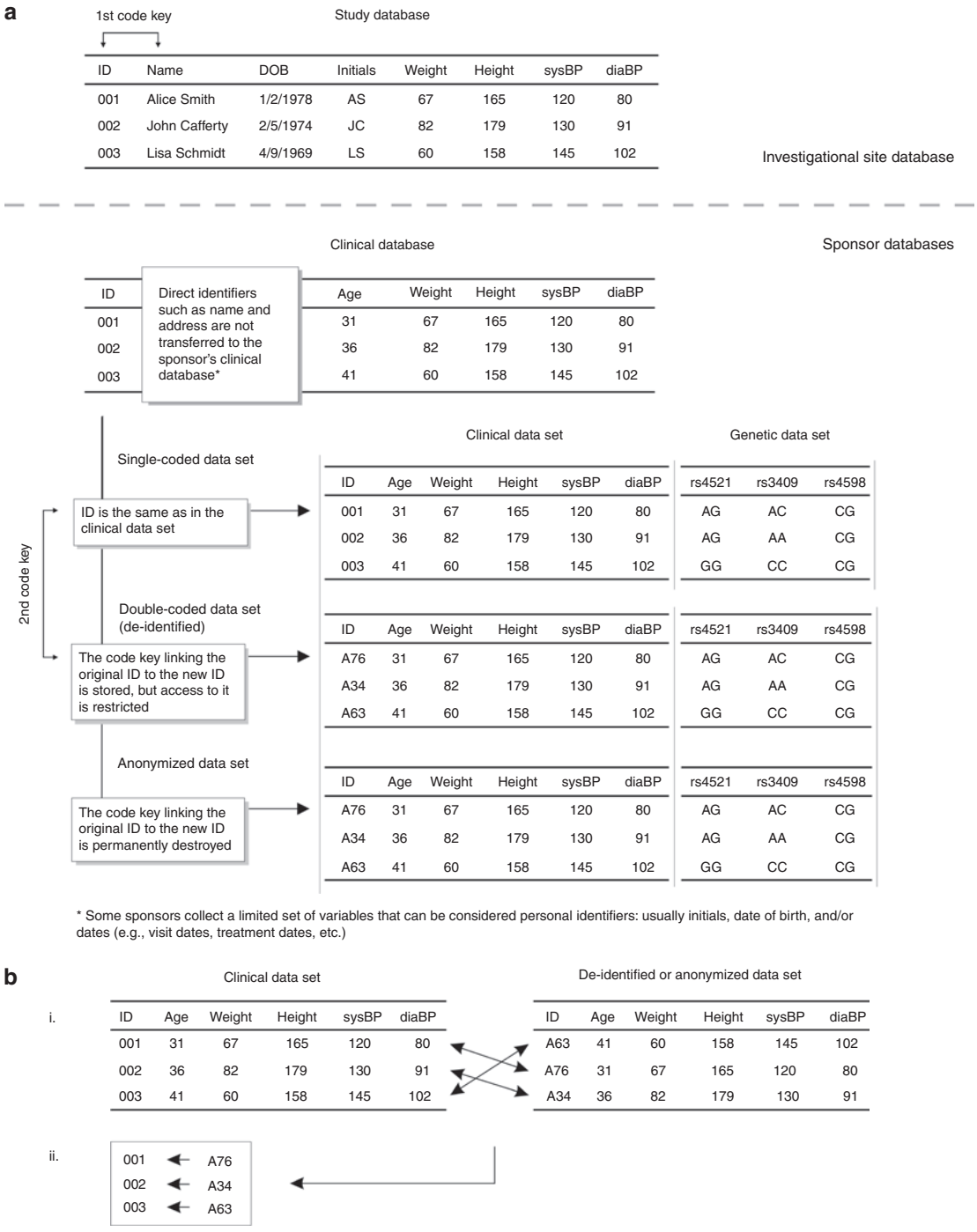

Figure 3 Coding methods and inherent limitations. (**a**) Sample and data coding methods applied in the pharmaceutical industry. The investigator is responsible for protecting the primary key that links the subject identifier to the subject's name. Direct personal identifiers (e.g., name, address, Social Security number) are not transferred to the sponsor's clinical database, although some sponsors continue to use initials, date of birth (DOB), and various forms of dates. Single-coded samples/data usually do not carry personal identifiers and can be linked back to the subject by means of the primary key held by the investigator. Double-coded/de-identified samples/data do not carry any personal identifiers and are relabeled with a new code, which is linked to the original code via a secondary key. Access to and use of the secondary key may be restricted either by standard operating procedures or by physical access. Samples/data can be traced back very indirectly to the subject via the use of both coding keys. Anonymized samples/data are double-coded/de-identified, with deletion of the secondary key. (**b**) Data from the de-identified or anonymized data set can theoretically be linked back to the original subject identifier by matching the clinical variables of the de-identified or anonymized data set with those of the single-coded clinical data set. Even a partial clinical data set for a subject effectively serves as a unique identifier for that subject by acting as a "clinical bar code" or "clinical fingerprint." Even after the unique original subject identifier is dissociated from the subject during de-identification/anonymization, the clinical bar code continues to effectively serve as a unique subject identifier. This clinical bar code can be used to reconstruct the link to the subject's original identifier. As a result, genomic data can be readily linked to the subject's original identifier by comparing the de-identified/anonymized clinical data with clinical data that contain the original identifier. The link back to the specific subject's name would however still require access to the primary key held by the investigator. diaBP, diastolic blood pressure; ID, identification; sysBP, systolic blood pressure.

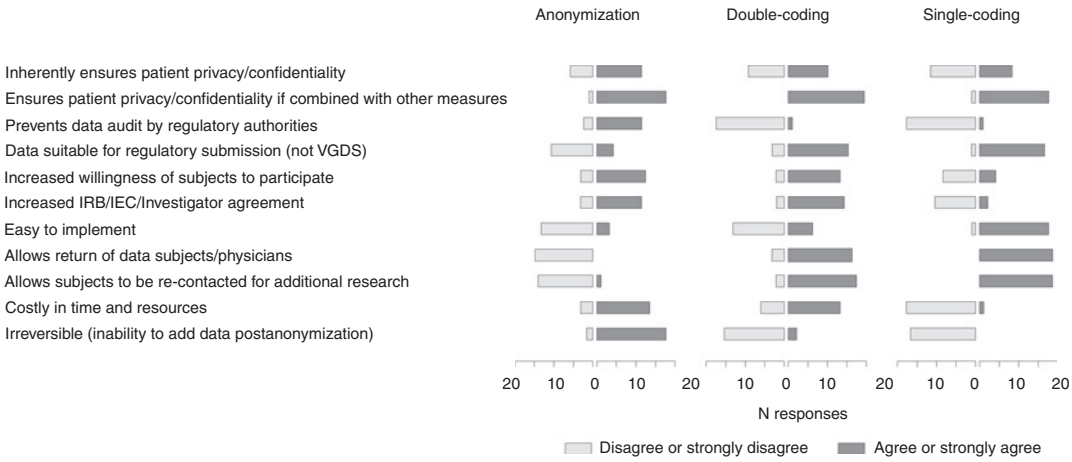

Figure 4 Industry's perceptions of the properties of coding methods. IEC, independent ethics committee; IRB, institutional review board; VGDS, voluntary genomic data submission.

of future research, avoiding the communication of genetic data to subjects in situations where the clinical validity of the findings is not established, omitting genetic research data from subjects' medical files, and ensuring overall compliance with good clinical practices, documentation, and SOPs. Many of these measures are also currently used for nongenetic samples and data.

Risks associated with storage for future research

Genetic samples stored long-term for future research have historically been perceived to pose a higher risk of privacy violation, largely owing to unknown factors such as how the information generated from the sample might be used, the impact of sophisticated technologies of the future, and improvements in data interpretation over time. The survey results describe measures that companies have implemented to minimize these risks. These include limiting the duration of storage, limiting the scope of research, maintaining samples in a format that allows subjects to withdraw consent in the future, and enforcing processes to maintain patient confidentiality. The more complex coding procedures are comparatively ineffective at minimizing potential risks associated with long-term storage.

Industry's track record in maintaining privacy of genetic information

To our knowledge, there are no publicized examples of pharmaceutical companies inadvertently disclosing genetic data with negative consequences to study participants. The industry has a good track record of maintaining tight controls and has highly regulated procedures in place for protecting patient privacy and confidentiality in research. The industry has been criticized for a certain unwillingness to contribute samples and data to research consortia; this is, in part, attributable to the very stringent procedures for privacy protection and commitments made in the informed-consent form.

Standards in pharmaceutical clinical trials

Pharmaceutical clinical trials are largely performed in such a way that the sponsor does not obtain the true identity (name, social security number, address, etc.) of the subject but uses a unique subject ID to relate information collected in the trial and analytical measurements used in patient monitoring or care. The key that associates the subject ID with an individual's identity is maintained by the independent investigator at the clinical site.[3] Although this key is accessible to certain sponsor representatives for quality assurance verification purposes, it is not otherwise available to the sponsor. This ensures data quality by maintaining an audit trail for data verification while maintaining subjects' privacy. Clinical trial databases are generally referred to as single-coded but will often contain a limited set of variables (generally initials, date of birth, and various forms of dates) that can be considered personal identifiers in certain countries. According to ICH E15, single-coded data and samples are "usually labeled with a single specific code and do not carry any personal identifiers," and single-coding "is the current standard used in clinical research and offers additional safeguards to the subject's identifiers as compared with the general health-care confidentiality and privacy protection in everyday medical practice" (**Tables 1** and **2**).

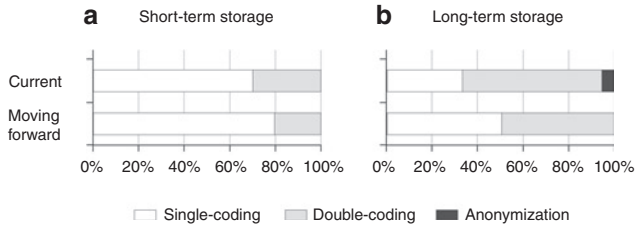

Figure 5 Recommendations by companies for the minimal level of coding required for DNA sample/data storage in the current global environment, and moving forward. Short-term storage was defined as up to 1 year after completion of a trial or until regulatory approval or until the compound under study is terminated. Long-term storage was defined as >1 year after the end of the trial or indefinitely.

Limitations of coding methods

The relationship between the coding methods applied to DNA samples and the keys linking subject codes to their identity is

illustrated in **Figure 3**. A limitation of single-coding for DNA samples is that IRBs/ECs currently may not broadly consider this an adequate standard for subjects' privacy protection even though this method is appropriate for other clinical research samples/data. In addition, regulations in some countries require more complex coding methods if the samples are intended for use in future research.[5]

Double-coded/de-identified samples/data do not contain personal identifiers and are relabeled with a new ID, and the key linking the original ID to the new ID (i.e., secondary key) is retained by the sponsor or designee (**Figure 3a**). The strength of this method hinges on (i) the conditions under which the secondary key is secured, (ii) the conditions under which it would be used to re-establish the link to the original ID, and (iii) restriction of access to both single-coded and double-coded/de-identified data sets. These conditions should be specified in an SOP/policy. Double-coding/de-identification is, by definition, nonpermanent, given that the secondary key is retained and not destroyed/deleted. This can be considered an advantage because it allows certain actions such as data communication to subjects when appropriate, submission to regulatory authorities, addition of data, and consent withdrawal. It may be considered a disadvantage, however, given that re-establishment of the link renders samples/data no different from single-coded samples/data. Importantly, double-coding/de-identification suffers from the same flaw that does anonymization (described below), namely, that it is susceptible to reconstruction of the link between the original ID and new ID through matching of clinical data parameters (illustrated in **Figure 3b**). The patient's identity is still ultimately protected by the primary key held by the investigator.

Anonymization carries the double-coding process one step further by permanently deleting the secondary key that links the original ID to the new ID (**Figure 3b**). This additional step potentially compromises these data for regulatory decision-making purposes and removes the ability to perform certain actions such as subject's withdrawal of consent. Furthermore, it is recognized that anonymization is susceptible to reconstruction of the link between the new ID and the original subject ID by means of cross-comparing anonymized data with single-coded data (**Figure 3b**) because the combination of clinical parameters serves as a clinical "bar code" that uniquely characterizes a study subject. Since single-coded data are generally broadly available to pharmaceutical company research personnel who have a legitimate need to access and use those data for research, it can be a challenge to prevent such relinking of IDs. A specific policy or SOP prohibiting reconstruction of the link between genetic data and the original subject ID is a useful complement to anonymization itself, but it is effective only to the extent that persons having access to anonymized data, in any form, are trained on the SOP. In the eventuality of link reconstruction, the anonymized data are no different from single-coded data. Ultimately, the primary key (held by the investigator) is the principal measure that protects the identity of the patient.

The consequences of anonymization described in ICH E15 (**Table 2**) may be interpreted to suggest that anonymization is a permanent, irreversible process that offers comprehensive privacy protection. However, anonymization is not without boundaries, and the true added value of this method remains questionable in light of the shortcomings described. Like de-identification, anonymization offers little added privacy protection while possibly creating a misleading sense of additional security.

Several companies indicated that they use statistical disclosure limitations methodology (although this is not a coding method per ICH E15) whereby data values are stored as ranges rather than individual values. This method suffers from heavy resource requirements and reduced utility of specimens to generate valuable research data, because the resolution is decreased.

Existing pharmaceutical standards for subject privacy

Maintaining the confidentiality of study participants and the privacy of their information is of paramount concern to industry researchers, regulators, and patients. Standard practices for data privacy protection in industry research have been in existence for a considerable length of time, and they meet regulatory requirements for patient privacy and appropriate conduct of clinical research. The industry operates minimally under the principles of good clinical practice, a standard that provides assurance that confidentiality of subjects is protected and in which confidentiality is defined as "the prevention of disclosure, to other than authorized individuals, of a sponsor's proprietary information or of a subject's identity." This standard dictates that "the confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with applicable regulatory requirements."[3] Therefore, a strong case can be made that adequate patient privacy requirements are already being fulfilled if these routine procedures are also applied to samples stored for future research.

Summary and future directions

Limited added value of complex coding methods in the context of industry. In the context of the pharmaceutical industry, sophisticated coding methods, by their inherent nature, offer limited incremental privacy protection while hindering productive research since its utility is affected by the amount and type of clinical data associated with a sample as well as the submissibility of such data to regulatory authorities and the resources required to double-code/de-identify or anonymize samples and data. These methods can also create a false sense of security if IRBs/ECs and study participants interpret them to be fail-safe. Ultimately, it is the separation of study-site investigators (who guard the primary key) from the sponsors that provides the primary protection of a subject's identity, whether for genetic data or other clinical trial data.

Personal identifiers. There is no universal standard for what constitutes "personal identifiers," and even those listed under HIPAA can be open to interpretation, notably "any other unique identifying number, characteristic, or code." ICH defines personal identifiers rather vaguely, and the definition of "personal data" in the European Union Directive is subject to interpretation by

individual Member States (**Table 3**). Not surprisingly, the survey indicated a wide range of definitions across the industry.

Of the parameters that can be considered personal identifiers, those generally applicable to the pharmaceutical industry are initials, date of birth, and dates directly related to an individual. Initials and date of birth are sometimes collected primarily for quality assurance purposes. Provided that substitute methods can be implemented to ensure accurate subject–data correspondence, the collection of these identifiers by the sponsor can be minimized. Some companies are already moving in this direction in order to increase privacy protection, facilitate global research, and circumvent the need for secondary coding methods. Importantly, however, dates directly related to an individual are critical to the conduct and analysis of clinical trials and cannot readily be eliminated. (In the United States, under HIPAA, a limited data set of personal identifiers is permitted for research purposes when a data-use agreement is in effect. The limited data set permits the use of some personal identifiers such as dates, city, state and zip codes; however, the effect of these identifiers in the limited data set is not fully examined in this paper.)

In view of the principle that international standards should not conflict with country-specific laws and regulations, it is understandable that ICH E15 does not provide a specific definition for "personal identifiers." In the United States, under the privacy rule, a data set is considered de-identified if the 18 personal identifiers are removed.[8] Since the privacy rule does not permit an individual to grant authorization for nonspecific future research,[14] samples should be de-identified; alternatively, either the subject's re-consent or an IRB/EC waiver may be obtained. HIPAA has come under criticism by the Institute of Medicine for provisions that hinder research while failing to provide substantive privacy protection.[11] There is an important need for a clear and practical definition of personal identifiers in the context of sample/data coding practices for stored samples in pharmaceutical research.

Coding practices for DNA. Survey respondents were evenly divided with respect to recommendations for coding practices for long-term DNA storage moving forward. Those who indicated a preference for single-coding were equal in number to those who preferred double-coding/de-identification. The survey did not request the reasoning underlying a company's recommendation, although free-text comments indicated a certain pragmatic resignation to accepting whichever process would allow DNA collection and storage in as many jurisdictions as possible.

Considering the well-established standard of single-coding in pharmaceutical research and the inherent limitations of the coding methods described here, the I-PWG recommends a dialogue with IRBs/ECs, health authorities and key stakeholders, in order to abrogate genetic exceptionalism with respect to coding practices such that equal standards are applied to DNA and non-DNA samples, whether for study-specific purposes or for future research. We believe that the standards applied in the industry provide adequate safeguards for privacy; however, we acknowledge that current regulations in certain jurisdictions pose obstacles to applying this standard to samples stored for future research. Our recognition of single-coding as an appropriate standard is motivated by (i) the established stringency of privacy standards applicable to industry (e.g., good clinical practice), (ii) an understanding of the limited value of complex coding methods, (iii) a need for more efficient data use and integration, and (iv) increased requirements by regulatory agencies to understand the contribution of genetics to drug response. We propose a dialogue with key stakeholders for more universal standards for samples stored for future research.

There is a clear need for improved understanding and harmonization, at the same time respecting cultural, legal, and other differences across jurisdictions. It would be advantageous to open communication across industry and nonindustry parties, particularly those considering stricter coding practices,[15] so as to ensure that relevant experience and expertise are shared toward standards that are acceptable to the scientific community and public at large. The pharmaceutical industry is committed to continuing to garner the trust of patients and the medical community by sustaining high standards for data privacy protection.

METHODS

The I-PWG (formerly the Pharmacogenetics Working Group) developed a questionnaire to solicit information about current DNA sample collection and data coding practices in the pharmaceutical industry. The I-PWG, a voluntary association of pharmaceutical companies engaged in pharmacogenomic research, focuses on noncompetitive topics. The full questionnaire included 54 questions, most of which were multiple-choice. Questions related to DNA coding practices (16 questions), including tallied responses, are presented in the **Supplementary Data** online. Results pertaining to data collection and storage practices are reported by Franc *et al.* in this issue.[1] All questions were optional. Percentages were calculated using the number of responses to the question as the denominator.

Invitations to participate in the survey were distributed in June 2009 to pharmacogenomics representatives of 31 pharmaceutical companies that were members of the I-PWG, the Pharmaceutical Research and Manufacturers of America, or both. The initial period for submission of responses was 40 days (it was extended to 60 days). Qualified representatives from each company involved in DNA-based research in clinical trials were asked to serve as primary responders and to seek collective group opinion from their colleagues, where possible.

The Web-based survey system (SurveyMonkey.com) that was used for data collection and analysis was set not to collect any information that could identify a particular company or representative. Participants were informed that the data would be analyzed and presented in aggregate such that their responses would be anonymous and that results would not associate to any particular practice or approach of a respondent company.

Responses were received from 21 companies (a 68% response rate). The companies that contributed data, as recorded by an independent legal party, were Abbott, Amgen, AstraZeneca, Bayer–Schering, Biogen Idec, Boehringer–Ingelheim, Bristol–Myers Squibb, Eisai, Genentech, GlaxoSmithKline, Johnson & Johnson Pharmaceutical Research & Development, Merck and Co. Inc., Millennium, Novartis, Orion Pharma, Pfizer, Roche, Schering Plough, Takeda, Teva Pharmaceuticals, and UCB.

SUPPLEMENTARY MATERIAL is linked to the online version of the paper at http://www.nature.com/cpt

ACKNOWLEDGMENTS

We acknowledge the contributions of Jenny Green in administrative assistance and of Charles Lister (Covington and Burling LLP) for legal monitoring.

CONFLICT OF INTEREST

All the authors are employed by pharmaceutical companies that are actively engaged in pharmacogenomic research: collecting, analyzing and storing DNA samples from subjects participating in clinical trials. This article does not necessarily reflect the views of the companies that are members of the I-PWG.

1. Franc, M.A., Warner, A.W., Cohen, N., Shaw, P.M., Groenen, P. & Snapir, A. Current practices for DNA sample collection and storage in the pharmaceutical industry, and potential areas for harmonization. *Clin. Pharmacol. Ther.* **89**, 546–553 (2011).
2. Harmon, A. Tribe wins fight over research on its DNA. *New York Times*, 22 April 2010.
3. International Conference on Harmonisation. ICH Guideline for Good Clinical Practice E6(R1) <http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf> (1996).
4. Ricci, D.S. *et al*. Global requirements for DNA sample collections: results of a survey of 204 ethics committees in 40 countries. *Clin. Pharmacol. Ther.* **89**, 554–561 (2011).
5. Warner, A.W. *et al*. Challenges in obtaining adequate genetic sample sets in clinical trials: the perspective of the Industry Pharmacogenomics Working Group. *Clin. Pharmacol. Ther.* **89**, 529–536 (2011).
6. International Conference on Harmonisation. ICH Guideline for Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories E15 <http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E15/Step4/E15_Guideline.pdf> (2007).
7. Spear, B.B. *et al*. Terminology for sample collection in clinical genetic studies. *Pharmacogenomics J.* **1**, 101–103 (2001).
8. Health Insurance Portability and Accountability Act (HIPAA) of 1996 (45CFR 164.514) <http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf>.
9. Directive 95/46/EC of the European Parliament and of the Council <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML> (1995).
10. European Commission. Data Protection Working Party. Opinion 4/2007 on the concept of personal data <http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf> (2007).
11. *Beyond the HIPAA Privacy Rule* (National Academies Press, Washington, DC, 2009).
12. Lin, Z., Owen, A.B. & Altman, R.B. Genetics. Genomic research and human subject privacy. *Science* **305**, 183 (2004).
13. Sweeney, L. Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics* **25**, 98–110, 82 (1997).
14. 67 Fed. Reg. 53181, 53226 (2002).
15. Office for Civil Rights. Workshop on the HIPAA Privacy Rule's De-Identification Standard. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/deidentificationworkshop2010.html>.