INTERNATIONAL
PHARMACEUTICAL
PRIVACY CONSORTIUM

IPPC White Paper on Anonymisation of Clinical Trial Data Sets

Overview

The data generated in the course of clinical trial research can be of great value to medical researchers, enabling additional research analyses to be conducted and increasing transparency around and confidence in reports of clinical trial outcomes. Health authorities in many countries around the world have encouraged pharmaceutical companies to share patient-level data collected during clinical trial research with third-party researchers. In response, the European Federation of Pharmaceutical Industries and Associations (EFPIA) and the Pharmaceutical Research and Manufacturers of America (PhRMA) have implemented a set of joint *Principles for Responsible Clinical Trial Data Sharing*.[1] In addition, policymakers in many countries and regions are considering approaches for making patient-level clinical trial data available to researchers.[2]

While many recognize the value of greater data access, concerns about patient privacy may limit the extent to which data can be provided to other researchers. The International Pharmaceutical Privacy Consortium (IPPC) believes that patient privacy can be appropriately protected by:

- removing identifying personal information from clinical trial data sets while still preserving the data's utility to researchers.

- providing "anonymised" data with accompanying contractual and organizational controls that, for example, prohibit data recipients from attempting to re-identify individuals and restricting access to only using the data within a secure environment.

A common understanding of what constitutes an "anonymised" data set in this context would help advance the development of policies and procedures for clinical trial data sharing. Accordingly, the IPPC presents the following framework for removing identifying personal information from clinical trial data sets. The IPPC urges regulators to consider data sets that have been processed according to this framework to be "anonymised" or "de-identified" data sets, as described further below. Data sets processed according to this framework will no longer contain information that can reasonably be used by a recipient to identify an individual clinical trial participant.

[1] *See* http://transparency.efpia.eu/uploads/Modules/Documents/data-sharing-prin-final.pdf.

[2] *See, e.g.,* European Medicines Agency, "Release of data from clinical trials," available at http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/general/general_content_000555.jsp&mid=WC0b01ac0580607bfa; US Food and Drug Administration, "Availability of Masked and De-identified Non-Summary Safety and Efficacy Data; Request for Comments" (June 4, 2013), available at https://www.federalregister.gov/articles/2013/06/04/2013-13083/availability-of-masked-and-de-identified-non-summary-safety-and-efficacy-data-request-for-comments.

Legal Background

Data privacy laws in many countries protect the confidentiality of patient health information. In some jurisdictions (e.g., the European Union), there are omnibus laws that apply to personal data generally. In others (e.g., the United States), there are sector-specific laws that apply to health information. These laws recognize a right of the individual to determine for what purposes his or her personal health information may be collected, used, and disclosed. In all jurisdictions, however, the need to balance informational autonomy with beneficial uses of data is accepted, and data is protected only insofar as it relates to an identified or identifiable individual.

In clinical trial research, "identified" data is held only by the researcher or research site conducting the trial. Rather than providing this data directly to a research sponsor, like a pharmaceutical company, researchers instead provide "key-coded" versions of this data. In "key-coded" data, some identifying information—like names, initials, dates of birth, etc.—has already been removed and replaced with an alphanumeric (or "key") code. Because key-coded data may still contain information capable of being used to identify an individual, it is protected under the EU's privacy directive.[3] Data that has been stripped of identifying elements is variously called "de-identified" or "anonymised." In this paper, we refer to such data that has been stripped of all identifiers as "anonymised" in order to distinguish partially de-identified "key-coded" data from fully de-identified data.[4]

[3] The EU Data Privacy Directive holds that personal data includes "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."

[4] The UK Information Commissioner's Office explains the key-coding process further:

> In a clinical study, only key-coded data is reported by clinical investigators (healthcare professionals) to the pharmaceutical companies sponsoring the research. No personal data is disclosed. The decryption keys are held at study sites by the clinical investigators, who are prohibited under obligations of good clinical practice and professional confidentiality from revealing research subject identities. The sponsors of the research may share the key-coded data with affiliates overseas, scientific collaborators, and health regulatory authorities around the world. In all cases, however, recipients of the data are bound by obligations of confidentiality and restrictions on re-use and re-identification, whether imposed by contract or required by law. Given these safeguards, the risk of re-identification of the key-coded data disclosed by a pharmaceutical sponsor to a third party under such obligations is extremely low.

UK Information Commissioner's Office, "Anonymisation: Managing Data Protection Risk Code of Practice" at Annex 2, p.66.

In explaining the concept of "anonymised data," European data protection authorities have commented that "putting in place the appropriate state-of-the-art technical and organizational measures to protect the data against identification may make the difference to consider that the persons are not identifiable, taking account of all the means likely reasonably to be used by the controller or by any other person to identify the individuals."[5]

> In [some] areas of research or of the same project, re-identification of the data subject may have been excluded in the design of protocols and procedure, for instance because there is no therapeutical aspects involved. For technical or other reasons, there may still be a way to find out to what persons correspond what clinical data, but the identification is not supposed or expected to take place under any circumstance, and appropriate technical measures (e.g. cryptographic, irreversible hashing) have been put in place to prevent that from happening. In this case, even if identification of certain data subjects may take place despite all those protocols and measures (due to unforeseeable circumstances such as accidental matching of qualities of the data subject that reveal his/her identity), the information processed by the original controller may not be considered to relate to identified or identifiable individuals taking account of all the means likely reasonably to be used by the controller or by any other person. Its processing may thus not be subject to the provisions of the Directive. A different matter is that for the new controller who has effectively gained access to the identifiable information, it will undoubtedly be considered to be "personal data".[6]

Both technical (e.g., removal of identifiers, security of database) and organizational (e.g., legal controls prohibiting re-identification attempts, such as a data use agreement) can be used to prevent data from reasonably being used to identify data subjects.

The US Health Insurance Portability and Accountability Act (HIPAA) of 1996 follows a similar approach, limiting both the uses and disclosures of "protected health information" (PHI). PHI includes all information that relates to the health of an individual and with respect to which there is a reasonable basis to believe the information can be used to identify the individual. Health information that does not identify an individual, and where there is no reasonable basis to believe that the information can be used to identify an individual, is deemed to be "de-identified."

Under HIPAA, there are two permissible methods for de-identification of PHI. The first involves removal of the following identifiers of the individual or of relatives, employers, or household members of the individual: (1) names; (2) all geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code and their equivalent geocodes, with certain limited exceptions; (3) all elements of dates, except year, for dates directly related to an

[5] Article 29 Data Protection Working Party, "Opinion 4/2007 on the Concept of Personal Data (WP 136)" (June 20, 2007), at p.17.

[6] Id. at p.20.

individual, including birth date, admission date and discharge date; (4) telephone numbers; (5) fax numbers; (6) e-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) URLs; (15) IP address numbers; (16) biometric identifiers, including finger and voice prints; (17) full face photographs and comparable images; and (18) any other unique identifying number, characteristic or code.[7] There must be no actual knowledge that the remaining information could be used in combination with other information to identify a data subject.

The second permissible method of de-identification involves obtaining a certification from a person with appropriate experience with generally accepted statistical and scientific principles of methods of de-identification that "the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information." The results and methods of this analysis must be documented by the statistician. In guidance issued in 2012, regulatory officials clarified that for purposes of rendering health information de-identified, no specific professional degree or certification is required, and relevant expertise may be gained through various routes of education and experience.[8]

IPPC Position

Review of the legal background demonstrates that the determination of whether or not a data set has been "anonymised" requires an examination of the following elements: technical and organizational measures, contractual obligations, and the redaction of potentially identifying information. In the framework outlined in the following section, the IPPC describes an approach that it believes will produce an "anonymised" data set within the meaning of the EU Data Privacy Directive and the US HIPAA regulations. Such data sets are suitable for sharing with outside researchers, in accordance with the terms, conditions, and procedures outlined below (and subject to appropriate technical and organisational controls to prevent re-identification).

____

I. **PURPOSE**. This document has been developed by the International Pharmaceutical Privacy Consortium to describe an approach to the removal of identifying personal information from a clinical trial data set before that data is shared with researchers. In addition to the procedures described below, consideration should be given to

[7] 45 C.F.R. § 164.514(b)(2).

[8] Office for Civil Rights, Department of Health and Human Services, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act Privacy Rule" (Nov. 26, 2012).

organisational and technical measures that prevent the combination of the processed data with non-anonymised data.

II. DEFINITIONS.

A. Personal Information. Any information relating to an “**Identifiable Natural Person**.”

B. Identifiable Natural Person. A person who can be “**Identified**,” directly or indirectly, by reference to an identification number or one or more factors specific to his or her physical, physiological, mental, economic, cultural or social identity.

C. Identified. Distinguished from a group of other persons in a way that permits linkage to other data sets containing information about the same person. A person may be identified even if not named.

III. AGREEMENTS. Researchers receiving data sets which have been processed according to this framework must sign agreements promising to protect patient privacy, including not attempting to re-identify individuals and to comply with terms of agreements around data security and agreed use. Researchers must also promise not to combine the data sets processed under this framework with other data sets, including a promise not to attempt to re-identify individuals.

IV. DATA THAT MUST BE REMOVED. Where it appears, the following data must always be removed or replaced as described. The IPPC notes that most clinical trials do not collect much of the information described below, and the listing of these categories of data is for completeness purposes only.

A. Names. Including both full and partial names, including initials.

B. Geographic Subdivisions. Including street addresses, cities, counties, states, legislative districts, and postal codes representing areas with populations below 20,000 persons.

1. Postal codes may be replaced with numbers which indicate the same level of relative geographic proximity indicated by a postal code. For example, if the data set contains US zip codes 20001, 20002, and 20003, those codes may be replaced with 51, 52, and 53 to indicate the geographic proximity of the data subjects.

C. Day and Month Values in Dates.

1. For events occurring outside of the clinical study time period, only years may be indicated for dates related to identifying events in a data subject's life. This includes birth dates, hospital admission dates, health-care practitioner visit dates, and dates of death.

   Where appropriate, a birth date may be replaced with the data subject's age at the time the information was gathered. However, all persons over the age of 89 must be grouped into a single category.

2. For events occurring during the clinical study, three approaches are acceptable:

   (i) dates may be removed;

   (ii) dates may be expressed as the number of days that passed from the data subject's enrollment in the clinical study; or

   (iii) dates may be replaced with "dummy dates," which do not correspond to the actual dates in the data set but which preserve the temporal relationship between events.

D. **Contact Information**. Including, but not limited to, telephone numbers, fax numbers, email addresses, websites, URLs, and screen names.

E. **Identifying Numbers**. Including Social Security numbers, national identifying numbers, account numbers, medical record numbers, health insurance numbers, certificate numbers, license numbers, vehicle identification numbers, license plate numbers, device serial codes, Internet Protocol addresses, and other numbers capable of identifying a single person or a small number of persons.

Clinical trial participant numbers should also be removed and replaced with a second set of identification numbers. As a best practice, any key linking the two sets of numbers should be destroyed.[9]

[9] The IPPC supports the conclusion of the Article 29 Working Party in Opinion 4/2007 on the concept of personal data, available at http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf. There, the Article 29 Working Party concluded that controllers who possess key-coded data, but are operating within a "specific scheme" in which "re-identification is explicitly excluded and appropriate technical measures have been taken in this respect," are not engaged in "processing personal data." The Article 29 Working Party continued on to note that:

> In other areas of research or of the same project, re-identification of the data subject may have been excluded in the design of protocols and procedure, for instance because there is no therapeutical aspects involved. For technical or other reasons, there may still be a way to find out to what persons correspond

F. **Biometric Identifiers**. Including finger prints, voice recordings, pictures of identifying marks, full-face images, or any other picture that depicts a sufficient area of the data-subject in sufficient detail to permit re-identification.

V. **DATA WHICH MAY BE REMOVED OR REPLACED**. Where they appear, it may be necessary to remove the following data fields which function as "quasi-identifiers." Again, the IPPC notes that clinical trials may not collect all of the types of data described below.

A. **Investigator Information**. Including site name, investigator identification, and investigator affiliation should be removed or replaced with a random number. Investigator site information may also be aggregated to a national or regional level. Where appropriate, a list of sites or investigators who participated in the study can be provided, so long as individual data subjects are not linked to particular sites or investigators.

B. **Socioeconomic Data**. Including name of employer, job title, occupation, income, education, and place of work. Specific information may be replaced with broad categories (for example, "post-secondary education" rather than the name of a specific educational institution).

C. **Household, Family Composition, and Pregnancy Information**. Including names of relatives. Information regarding the exact number of pregnancies may be replaced with ranges where appropriate (0 to 2, 2 to 4, etc.).

D. **Ethnicity**. If the population associated with the data set is such that including ethnicity would create a risk of re-identification, ethnicity should be removed or replaced with the Clinical Data Interchange Standards Consortium's (CDISC) standard ethnicities.

E. **Adverse Events**. Adverse event descriptions or codes should be presented in a generalized manner that does not permit re-identification of the data subject.

what clinical data, but the identification is not supposed or expected to take place under any circumstance, and appropriate technical measures (e.g. cryptographic, irreversible hashing) have been put in place to prevent that from happening. In this case, even if identification of certain data subjects may take place despite all those protocols and measures (due to unforeseeable circumstances such as accidental matching of qualities of the data subject that reveal his/her identity), the information processed by the original controller may not be considered to relate to identified or identifiable individuals taking account of all the means likely reasonably to be used by the controller or by any other person. Its processing may thus not be subject to the provisions of the Directive.

F. **Genetic Information**. If the quantity of genetic information contained in the record could be used to match a subsequent genetic sample from the same individual to the data profile, then genetic information should be removed. Even a small set of genetic information, when combined with other factors, may be sufficient to identify a data subject.

G. **Verbatim Quotes**. Verbatim statements should be removed if they contain information which could be used to re-identify the data subject.

H. **Medical History**. If the data subject's medical history contains information about the data subject, or the data subject's family, which could permit re-identification, then medical history should be removed or replaced with generic language (e.g., "family history of heart disease").

I. **Other Free Text Fields**. Although the information found in free text fields which could identify a data subject generally falls into one of the categories described above, the deletion of other free text fields is generally recognized as an appropriate safeguard to prevent the accidental inclusion of identifying or quasi-identifying personal information.

VI. PROCEDURES AND TESTING.

A. **Procedure**. The process described in this framework should be carried out before the data set is provided to an outside researcher. As a best practice, any documents linking the researcher's data set to the original, identified data set should be destroyed.[10]

B. **Review**. Any data set must be reviewed for compliance with this framework before being released to an outside researcher. This review may be done by appropriate experts within the company charged with review of data sets for compliance with this framework, or it may be done by an outside expert bound by appropriate confidentiality obligations.

This review should also include a review of the data set to ensure that the remaining data does not contain unique or unusual information which may have the effect of indirectly identifying the data subject.

C. **Appropriate Documentation**. The process used to remove identifying personal information from any data set and the review of such data sets must be appropriately documented. The documentation should explain the purpose for which the data set was processed and to whom the data set will be released.

[10] See fn 9.