Final advice to the European Medicines Agency from the clinical trial advisory group on Protecting Patient Confidentiality

26 April 2013

1. Summary

Different types of data pose different levels of risk of identifying patients. In most cases, the risk is considered sufficiently low in case of reports containing only aggregated data, such as the main body of the clinical study reports and appendixes (excluding line-listings). After de-identification of any indirect identifiers (e.g., case narratives, outliers, tables with sparse numbers), such reports could be considered for publication and given unrestricted access. Applicant companies may use different transformation methods to de-identify the data. A recommended minimum standard for de-identifying data is described in Hrynaszkiewicz *et al.* (1).

In case of the raw data and line listings, the risk is considered much higher due to the combined presence of many indirect identifiers. Further de-identifying key-coded raw data and line listings is a resource-intensive task and may (debatably, in some or most cases) compromise the analytical validity of the data, so that unrestricted publication of fully de-identified raw data and line listings may not (debatably, in some or most cases) be useful. Access to the raw data and line listings (debatably, original key-coded or de-identified) should be allowed under similar rules to those applicable to processing of personal data by health care professionals subject to the obligation of professional secrecy. There should also be rules to ensure that additional use of raw data is within the scope of the informed consent/assent signed by trial subjects or on their behalf.

2. Problem statement

How can EMA ensure through its policy that patient and other personal information will be adequately protected i.e., that patients cannot be retroactively identified when clinical trial data are released, and that applicable legislation, standards, and rules regarding personal data protection will be respected?

3. Scope and definitions

3.1. This advice refers to any information containing clinical data (e.g., raw data, clinical study reports) that are submitted to the Agency as part of a marketing authorisation application, or subsequent submission (e.g., in the context of clinical variations of the marketing authorisation, submission of results of post-authorisation safety studies). When discussing the various options, a distinction is made between documents containing mainly aggregated data (i.e., the main body of the clinical study report and appendixes, excluding line-listings), and raw data and line listings, containing key-coded, patient-level data.

3.2. Personal data: 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 (2).
In this document, a distinction is made between persons included in clinical trials (e.g., patients or healthy volunteers and their legal representatives, hereinafter referred to as “subjects”), and any other person mentioned in the submission (investigators, study site personnel, sponsor representatives, contracted workers, etc., hereinafter referred as “clinical trial personnel”).

3.3. De-identified data: Data that have been made anonymous in such a way that the data subject is no longer identifiable (directly or indirectly).

3.4. Key-coded data: These data refer to information that relates to individuals that are assigned a code, while the key making the correspondence between the code and the common identifiers of the individuals (like name, date of birth, address) is kept separately. In clinical trials, the key is typically held by the investigators. Information to the pharmaceutical company or other parties involved is provided only in this coded form.

Key-coded data constitutes information relating to identifiable natural persons for all parties that might be involved in the possible identification and should be subject to the rules of data protection legislation (3).

4. **Clinical Trial Personnel's Data**

4.1. Option 1

Personal data of clinical trial personnel (name, CV, affiliation, etc.) are considered as professional information that is essential to be made public. Clinical trial personnel have legally defined responsibilities and roles with respect to aspects of the marketing authorisation dossier and the clinical trials that are part of the dossier. Assessment of the qualifications of the researchers and other clinical trial personnel is an important public interest in the area of public health protection and scientific research. Companies are advised that non-essential information (e.g. personal address, personal phone number) should not be included in the dossier.

Option 2

Personal data relating to the principal investigator and the experts who sign the clinical study report are considered as professional information that is essential to be made public. This is justified by grounds of important public interest in the area of public health protection and scientific research. For any other clinical trial personnel there is no presumption of important public interest why such data should be made public.

Option 3

There is no presumption of important public interest why any personal data of clinical trial personnel should be made public.

4.2. There should be sufficient protection for the privacy of pharmaceutical company employees and researchers that perform non-clinical research. Similar considerations should apply to personnel participating in research that could be considered to be sensitive or controversial. In such cases, companies should be allowed to justify de-identification of data related to clinical trial personnel.

Comments for Option 1

In general, for clinical trials there is no great concern for revealing the names of investigators and study/company personnel, as shown by the ample information generally in the public domain about the investigators involved (e.g., as listed as authors or investigators in publications of medical journals, including their affiliations, contact details and emails). In multinational studies it is also important to know who the investigator in charge in that country is.

It is important that detailed information about the ethics committees is made available. However disclosure of Ethics Committee Member names would be highly problematic because it would almost certainly deter people from serving on such research ethics committees. Members of ethics committees

and IRBs who are involved in approving highly controversial studies will not want to face a backlash from persons, groups, or companies that may be opposed to a particular study.

Comments for Options 2-3

Except for a few people (the principal investigator, the persons responsible for the study or its interpretation, the experts who sign the report), there is no public health interest for disclosing such information about any other clinical trial personnel or persons whose names may appear in the dossier. Data related to such persons should be considered as personal data, not to be released without adequate de-identification. There is also a concern that publishing all investigators' names may add to the risk of identifying the clinical trial subjects.

The assumption should be that such information should not be disclosed unless the relevant individual has consented, or unless the information becomes public as a result of publication of study data. Disclosing such information is contrary to the position taken by companies and the EMA when releasing information in other contexts. Releasing the names of such company employees can expose them to personal risks, particularly where the research involves technologies or techniques that some may find more controversial, such as stem cell or gene therapies.

5. **Subjects' Data**

5.1. Currently, subjects' clinical data are submitted as de-identified data (e.g., aggregated data in the form of tables within a clinical study report) or as key-coded data (e.g., using the subject identification code instead of the subject's name as part of line listings).

5.2. Apart from direct identification, there is a risk that clinical trial data may allow identifying the subjects indirectly, through a combination of potential indirect identifiers. For instance, a person may be identified indirectly by initials, date of birth, a telephone number, a car registration number, a social security number, a passport number or by a combination of significant criteria which allows him to be recognized by narrowing down the group to which he belongs (age, occupation, place of residence, etc.).

5.3. For all the clinical trial data to be submitted or requested by the Agency (e.g., study report, data set), including any subsequent revisions, the applicant company shall assess the risk of compromising subjects' identity in case of wide publication of those data. Assessment of the risk should take into particular consideration data that could be considered to be sensitive or controversial and that might lead to discrimination if the subject can be identified, as well as situations with an intrinsic higher risk of identification such as rare diseases.

If for any data the risk of compromising subjects' identity in case of wide publication of those data is considered to be absent or sufficiently low, the applicant company shall clearly label the data as "SUITABLE FOR PUBLICATION". If the risk is not considered sufficiently low, for study reports and, where applicable, for raw data (depending on the option chosen under 5.6), the company shall submit two sets of documents, the original documents clearly labelled as "NOT FOR PUBLICATION", and the documents containing de-identified data clearly labelled as "SUITABLE FOR PUBLICATION".

5.4. Metadata about the study reports and data sets should be provided so those wishing to seek data for reuse can easily see the nature of what is available. For example, datasets could be listed with: Name of the trial; intervention (drug/device) being studied; name(s) of investigator(s); number of subjects; file type(s) and format(s); sponsor of study; date(s) trial

conducted; date of submission of data to the EMA; trial registration number (EUDRACT; ISRCTN/NCT number).

5.5. Study reports: In most cases, aggregate statistics (frequencies, sums, etc., as found in the main body of the clinical study reports and appendixes excluding line listings) might be considered as sufficiently de-identified so as not to constitute personal data.

Companies should be encouraged not to include promotional material in the study reports. Adequate disclaimers and visual prompts (for example, watermarking the pages of the study report) should be considered to avoid that any information made widely available could be misunderstood as representing Agency views. The reader should be directed to contact a patient organisation or an expert (e.g. their own doctor) in case one needs to discuss the results. A link to relevant patients' organisations could be provided along each study report/raw data.

5.6. Raw data and line listings:

Option 1
In case of raw data and line-listings, adequately de-identified data can be valuable and de-identifying the data does not necessarily compromise the analytical utility of the data (4). Adequately de-identified data should be made available for wide access. The data to be made available may include all the data sets or a relevant subset (e.g., the main analysis set, containing a limited number of indirect identifiers so that the risk of compromising subjects' identity in case of wide publication of those data is considered to be absent or sufficiently low whilst preserving the ability to replicate the main analysis).

Option 2
In case of raw data and line-listings, with few exceptions, available methods for de-identifying personal data cannot achieve sufficient de-identification while preserving sufficient analytical utility of the data, particularly for safety data and case narratives for adverse events. Publication of such data would either fail to protect patient confidentiality or result in a burdensome yet futile exercise of no analytical use. Access to the original key-coded data should only be allowed under strict rules to ensure confidentiality and alignment of the purpose of access to the subjects' informed consent/assent. Such rules should be similar to those applicable to processing of personal data for the purposes of preventive medicine, medical diagnosis, the provision of care or treatment or the management of health-care services, and where those data are processed by a health professional subject under national law or rules established by national competent bodies to the obligation of professional secrecy or by another person also subject to an equivalent obligation of secrecy (3).
According to one view, even when accessed under such strict rules in Option 2, the data may have to be also fully de-identified.

5.7. Guidance should be provided in the format of standard templates for subjects' informed consent/assent to better inform subjects of possible further uses of the data in the interest of public health and under the chosen rules of engagement.

6. De-identification of personal data

6.1. Applicant companies may use different transformation methods to de-identify the data. Generally, using such methods, it is possible to adequately de-identify data in such a way that, taking into account all the means likely reasonably to be used to identify subjects, the risk of

identifying a subject does not exist or is negligible; such de-identified data are no longer considered as “personal data”.

6.2. A recommended minimum standard for de-identifying data is described in Hrynaszkiewicz (1). In some situations, this minimum standard should be supplemented by additional de-identification methods (e.g., statistical). The methods of de-identification should also be such that adherence will preclude subject de-identification even when applying linkages with other data carriers (e.g., social media).

6.3. De-identification methods shall be individually tailored to the specific dataset and situation to ensure that a maximum of information is available while at the same time ensuring sufficient personal data protection. Methods and extent of de-identification should be adapted to sensitive or controversial situations that might lead to discrimination if the subject can be identified, as well as situations with an intrinsic higher risk of identification such as rare diseases.

6.4. Applicant companies shall describe in general terms and justify, if appropriate, for each document the de-identification methods used. If the de-identification methods are deemed insufficient or excessive, the Agency shall ask the applicant company to further justify and if necessary modify the de-identification method.

6.5. The Agency should establish whether it wishes to systematically verify that the data submitted as de-identified data contain no personal data, or if this is considered the responsibility of the applicant company.

6.6. The Agency shall produce further guidance on the standards and methods for de-identifying data. Upon request, the Agency shall provide advice to applicant companies, (where necessary involving relevant patient groups and members of the public), on the adequacy of the methods for de-identifying data.

7. References

(1) Hrynaszkiewicz, I., M. L. Norton, et al. (2010). "Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers." BMJ **340**: c181.

(2) Art. 2 (a) of the Directive 95/46/EC.

(3) Opinion 4/2007 on the concept of personal data of the Article 29 Data Protection Working Party.

(4) Sandercock, P. A., M. Niewada, et al. (2011). "The International Stroke Trial database." Trials **12**: 101.