Draft advice to the European Medicines Agency from the clinical trial advisory group on Clinical trial data formats

The clinical trial advisory group at their TC on 7th March 2013 provided advice as follows:

1. The following definitions were agreed

1.1 This advice refers to all data recorded in a clinical trial (as part of documents and data aggregated or at patient level) that can be stored electronically and associated documentation (additional information that identifies and characterises the data properties such as dataset keys, variable definition, terminology, code lists) that is submitted by Applicants to the Agency throughout the life-cycle of medicinal products. The policy will be applied prospectively for future submissions to the Agency; it may include old clinical trial data.

1.2 Data formats, in this advice, refer to the organisation of information according to pre-set specifications that facilitate the storage, exchange, access, comprehension, analysis and archive of clinical data. It includes both the type of electronic files and the structure of the files, as well as associated documentation.

The data and associated documentation concerned by this policy may or may not be sourced via electronic tools (e.g., paper or electronic case report forms), and are subsequently submitted and stored electronically.

All statements in this advice are made in the consideration that CTAG1 rules for patient data confidentiality and anonymisation are applied and effective, and that CTAG5 legal rules are strictly followed. As a consequence, there will be no more reference to the CTAG1 and CTAG5 rules in the rest of this advice.

2. There is a need to define data formats

The choice of formats should neither imply delays in the information to be made available nor impose unnecessary burden to the stakeholders.

Formats may be different depending on the type of information to be made publicly available and the intended use of it, and data should be made available irrespectively.

As there are not universally agreed formats, a minimum set of rules should be defined, including:

- An indexed list of all clinical trials present in the submissions shall be provided (if not already available in the table of content of the submission dossier) so the data of the overall clinical program is tracked. In this list, clinical trials should ideally be identified by a unique trial identifier. This identifier could be either: the EudraCT number (but it would only cover trials conducted in Europe and it is not commonly referenced in journals and published articles), the NIH clinicaltrials.gov website registry number (commonly referred to in the literature), the ISRCTN registry number or a number provided by the applicant at the time of submission. It is thought to be useful to be able to link back clinical trials to journal article information.

- Data shall be published in the format they have been submitted and evaluated and no conversion of formats will be done by either the marketing authorisation holder or the European Medicines Agency (EMA).
- Consistency of formats throughout the life cycle of the medicinal products is not mandatory but should be sought when achievable, e.g. for contemporaneous studies.
- Documents containing data should be human readable and searchable by anyone requesting the data from the EMA.
- There was a request that analysis of patient-level data could be done in Excel.
- Patient-level data should be accompanied by associated documentation that allows quickly grasping the data and processing it. This documentation, which includes metadata (= 'structured data about data'), should ideally be machine readable. For example, the documentation explains the structure of the data (e.g., what information is contained in each dataset), gives the definition of data elements (e.g., '1' corresponds to 'male' and '2' to 'female'), and provides the context to interpret correctly the data, to allow further analyses, without needing for additional information from neither the marketing authorisation holder nor the EMA.
- Formats should be chosen so that data is readable with open source, non-proprietary software (but not necessarily free): that includes, but is not limited to, portable document format (PDF) for text documents such as clinical study reports, SAS transport file format (XPT) for datasets and programs (as opposed to SAS format which is proprietary), and extensible markup language (XML) format for associated documentation on data. It would be easier for Industry in general if these requirements are the same as FDA's, although it is not favoured by small- and medium-sized enterprise if at a non-negligible cost.

3. What is to be included in data formats

Assuming that data privacy protection has been ensured for all data made available publicly, cCertain information such as CT scans, MRI and other imaging, interviews, genetic/genomic data can bring useful information and should be in the scope of discussion for data formats. However, that particular type of data is contained in large files; thus its transport, storage and access might cause serious informatics problems.

Three levels of clinical trial information, data and associated documentation shall be included.

- <u>Level 1</u>: for each product, a full list of clinical trials, including a unique study identifier; these lists should be fully searchable and could be connected to the European Public Assessment Reports. This is separate to information stored in the EUdraCT database.
- <u>Level 2</u>: for each study, full clinical study report (CSR) according to ICH E3, including all appendices, as detailed in ICH E3 (study information, patient data listings and case report forms [CRF]).
- <u>Level 3</u>: for each study, individual patient data sets (including individual patient data) and additional results used for the evaluation of the drug (if not covered by Level 2), documentation explaining the structure and content of datasets (e.g., annotated CRF, variable definitions, data derivation specifications, dataset define file), test outputs, SAS logs and SAS programs

Elements included in the three levels of data listed above may need to be modified in special circumstances driven by confidentiality or legal aspects.

4. Formats recommended

In general, to avoid delays any format shall be acceptable for all data until the policy is applied by stakeholders. The data shall be published in the format they are available at present.

In terms of the different types of data described in the previous section, Level 1 data should be searchable. PDF is recommended.

For Level 2 data (CSR and appendices, according to ICH E3), it should also be searchable. PDF is recommended. Of note, old CSRs may not fully comply with the current ICH E3 format. In this case, it will be acceptable to provide the CSR in the original format in which it was written.

Individual patient data and associated documentation (Level 3) shall be published in the format they are available at time of submission. That can be according to CDISC standards, and there was general agreement that Applicants will move progressively to an increase use of CDISC standards.

It was recognised that CDISC have defined useful formats: SDTM for data tabulations, ADaM for analysis datasets, and define xml for metadata. The recommendation is for all these to be submitted to the Agency, but not ODM, which is a transport format for data management. SDTM-annotated CRF would also be very useful for data re-analysis. It was acknowledged that CDISC implementation guides can be interpreted in different ways by Applicants, therefore EMA should define clear requirements in relation to these guides.

If other formats can be used, EMA should define minimal requirements of more basic formats, such as the following: clinical data should be submitted in rectangular tables, in a comma-separated values (CSV) format; associated metadata should contain at least one table with all datasets, all variables and their meanings, also possibly associated code lists, and another table with all codes and decodes, and the variables they relate to.

Individual data such as CRF data in PDF format are not useful as they will require substantial manpower for reloading in another usable format). However, PDF scans of printed out CRFs might be the minimal standard which is realisable even in a small academic institution or a small- and medium-sized enterprise, in order not to add unnecessary financial and resource burden to the marketing authorisation holder. The general view is that re-formatting of old data should be not requested by EMA; however, some are of the opinion that EMA should ask the marketing authorisation holder to provide the data in a format which is machine-readable and can be done with a non-proprietary software.

Harmonisation of formats such as CDISC SDTM and ADAM is of course desirable as this expands the usefulness of the data made available. This exercise shall be progressively implemented in a collaborative way between CDISC and EMA to ensure consistency and versioning control.

Sustainability of a chosen standard might also require reducing the speed of versioning and ensuring availability of software adapted to the subsequent changes of the formats. EMA guidance on formats may not follow the evolution of CDISC modifications at the same rhythm if it imposes too much burden on applicants. This will reduce the potential for re-formatting should a newer version be required.

Formats used across a number of studies for the same product do not need be compatible, although it will be a bonus when it can be achieved. For the datasets there is a need to:

- Harmonise a reference format worldwide
- Maintain versioning over time

A point to discuss further concerns mixed formats acceptability, e.g. for fixed combination of old and new active substances or hybrid mixed submission, when both clinical data from old studies and from new clinical trials are included.

5. Who should adhere to the agreed formats

The formats agreed are to be adhered to by all stakeholders and also for locally run clinical trials outside Europe if they become part of a submission to EMA. The Applicants should ensure correct implementation of the formats and should also consider implication of terms translations from different languages.

For clinical trials owned in different measure by multiple partners (e.g. public-private partnerships), the above points should be taken into account from the beginning of the clinical studies. This concerns data that are part of studies that are submitted to the Agency and where the marketing authorisation holder is legally permitted to share the data.

6. Timelines for format implementation

- While it seems reasonable to gain experience with formats of individual patient data (Level 3), it is not recommended to have a test period for clinical study reports, because the format of the CSRs, i.e. ICH E3, is in effect since 1996. Therefore the format for CSRs (Level 2) - and for Level 1 - can be mandatory from the implementation of the policy.
- Pro-active adoption of standard formats for Level 3 data: as this has to be mandatory for the sake of fairness and clarity for all stakeholders, it is advised to start gradually to acquire experience and then mandate formats after a trial period for all new studies submitted.
- At the end of this trial period, all levels of data can be released at the same time.

7. International harmonisation across regulatory agencies

The EMA is leading in terms of policy but global alignment and harmonisation are critical steps in the future process. A global consultation of formats is recommended at the ICH level (for human products and at the VICH level for veterinary products). The list of elements discussed in Section 3 and the corresponding formats discussed in Section 4 need to be included in that consultation. Communication with other national medicines agencies would also be beneficial. The policy should also aim at implementing what will be widely used in future to further standardise the process and prevent any re-formatting.

Under e-CTD, PDF, XML and other standards are allowed in MAA. ISO, CEN and CDISC to define CSRs harmonised standards.

8. References

ICH E3 Structure and Content of Clinical Study Reports

http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E3/E3_Guideline.pdf

Clinical Data Interchange Standards Committee

http://www.cdisc.org

Annex I - Comments from participants below may or may not have been made on behalf of the organisation they are affiliated with.

Line number	Comment and Changes proposed	Name	Affiliation
61	Comment: The title could be made clearer by making reference to "types" or "elements" of data, as this is the key section where the specific types of data are specified and divided into three levels (Level 1, 2, 3). **Proposed change (if any)**: Rename section as "What types of data (and in what format) are to be included".	Peter Doshi	Johns Hopkins University School of Medicine

Line number	Comment and Changes proposed	Name	Affiliation
61	Comment: The clinical study report (CSR, defined by ICH E3) and SAS datasets are two important types of trial data, but they are not the only types of trial data. As a major function of CTAG2 is to define the types of clinical trial data that--if suitably de-identified--can be prospectively released, it is important to include within the CTAG2 advice document a clear table that outlines the many types of data relevant to understanding and interpreting clinical trials. For each type of data that exists, EMA may already routinely be requesting these data from marketing authorization applicants, or they may consider requesting these data in the future. Whether EMA requests or intends to routinely request these data will clearly impact the ability of EMA to prospectively release these data. Those writing the EMA's draft policy should know whether EMA plans to hold such information itself. Part of the feasibility of prospective data release also depends on the format of the data (e.g. paper, PDF, or proprietary computer file format). Those writing the EMA's draft policy should have this information in order to better decide the feasibility and mechanism by which the data may be de-identified and released. Different de-identification issues may arise with different types of data. For instance, a trial's statistical analysis plan (ICH E3 section 16.1.9) has no information about patients, and presumably does not require de-identification before prospective release, whereas the list of study investigators (ICH E3 section 16.1.4) and a patient-level SAS dataset may require de-identification prior to release. There is a need to record the need for CTAG1 advice on de-identification for the types of data that CTAG2 identifies.	Peter Doshi	Johns Hopkins University School of Medicine

Line number	Comment and Changes proposed	Name	Affiliation
	Proposed change (if any): I propose adding a table to section 3 titled "Table of types of data" to this section that consists of 6 columns and accomplishes the above described needs. The columns would be as follows: Column 1: Type of data Column 2: Routinely requested by EMA (yes/no)? Column 3: Format (e.g. paper, PDF, electronic dataset such as SAS XPORT, etc.) Column 4: Level of data (i.e. Level 1, 2, or 3) Column 5: Need for CTAG1 advice on risk of participant re-identification (yes/no)? Column 6: CTAG1's advice on who, when (pre-submission or post-submission to EMA), and how to ensure acceptably low risk of re-identification (leave blank until CTAG1's advice is received) ================================ For Column 1 (type of data), I suggest it be populated with the following: - For each product, a full list of clinical trials, including a unique study identifier, the study title, the interventions and the indication studies; these lists should be fully searchable. The studies should be connected to the European Public Assessment Reports, to EUDRACT and to clinicaltrials.gov. [LEVEL 1] - Full Clinical Study Report (CSR) including all appendices defined by ICH E3 (http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002832.pdf) [LEVEL 2] - ICH E3 1-15 Core report - ICH E3 16.1 Study information - ICH E3 16.1.1 Protocol and protocol amendments	Peter Doshi	Johns Hopkins University School of Medicine

Line number	Comment and Changes proposed	Name	Affiliation
	- ICH E3 16.1.2 Sample case report form (unique pages only) - ICH E3 16.1.3 List of IECs or IRBs (plus the name of the committee Chair if required by the regulatory authority) - Representative written information for patient and sample consent forms - ICH E3 16.1.4 List and description of investigators and other important participants in the study, including brief (1 page) CVs or equivalent summaries of training and experience relevant to the performance of the clinical study - ICH E3 16.1.5 Signatures of principal or coordinating investigator(s) or sponsor's responsible medical officer, depending on the regulatory authority's requirement - ICH E3 16.1.6 Listing of patients receiving test drug(s)/investigational product(s) from specific batches, where more than one batch was used - ICH E3 16.1.7 Randomisation scheme and codes (patient identification and treatment assigned - ICH E3 16.1.8 Audit certificates (if available) - ICH E3 16.1.9 Documentation of statistical methods - ICH E3 16.1.10 Documentation of inter-laboratory standardisation methods and quality assurance procedures if used - ICH E3 16.1.11 Publications based on the study - ICH E3 16.1.12 Important publications referenced in the report - ICH E3 16.2 Patient data listings - ICH E3 16.3 Case Report Forms - Other documents containing contextual information not identified in ICH E3: - certificate(s) of analysis [LEVEL 2] - investigator's brochure [LEVEL 2] - manual of operations and procedures [LEVEL 2] - annotated CRFs [LEVEL 3]		

Line number	Comment and Changes proposed	Name	Affiliation
	- Electronic database of Individual Participant Data (IPD) - patient-level dataset (raw and derived) [LEVEL 3] - analysis datasets [LEVEL 3] - Contextual information to understand and work with electronic database of IPD - dataset specifications (metadata which describes the variable labels, variable descriptions, code lists and formats) - SAS programs [LEVEL 3] - SAS logs [LEVEL 3] - test outputs [LEVEL 2] - Original participant level records - filled out (completed) CRFs for all trial participants [LEVEL 3] - laboratory reports for all trial participants [LEVEL 3] - medical records and diagnostic reports for all trial participants obtained as part of trial procedures [LEVEL 3] - Documents related to clinical trials often created by trial sponsors - Marketing Assessments - Email correspondence - Meeting minutes - Records of the Data Monitoring Committee (also known as DSMB) e.g. adjudication committee		

Line number	Comment and Changes proposed	Name	Affiliation
	- ICH E3 16.1.1 Protocol and protocol amendments - ICH E3 16.1.2 Sample case report form (unique pages only) - ICH E3 16.1.3 List of IECs or IRBs (plus the name of the committee Chair if required by the regulatory authority) - Representative written information for patient and sample consent forms - ICH E3 16.1.4 List and description of investigators and other important participants in the study, including brief (1 page) CVs or equivalent summaries of training and experience relevant to the performance of the clinical study - ICH E3 16.1.5 Signatures of principal or coordinating investigator(s) or sponsor's responsible medical officer, depending on the regulatory authority's requirement - ICH E3 16.1.6 Listing of patients receiving test drug(s)/investigational product(s) from specific batches, where more than one batch was used - ICH E3 16.1.7 Randomisation scheme and codes (patient identification and treatment assigned - ICH E3 16.1.8 Audit certificates (if available) - ICH E3 16.1.9 Documentation of statistical methods - ICH E3 16.1.10 Documentation of inter-laboratory standardisation methods and quality assurance procedures if used - ICH E3 16.1.11 Publications based on the study - ICH E3 16.1.12 Important publications referenced in the report - ICH E3 16.2 Patient data listings - ICH E3 16.3 Case Report Forms	Peter Doshi	Johns Hopkins University School of Medicine

Line number	Comment and Changes proposed	Name	Affiliation
	- Other documents containing contextual information not identified in ICH E3: - certificate(s) of analysis [LEVEL 2] - investigator's brochure [LEVEL 2] - manual of operations and procedures [LEVEL 2] - annotated CRFs [LEVEL 3] - Electronic database of Individual Participant Data (IPD) - patient-level dataset (raw and derived) [LEVEL 3] - analysis datasets [LEVEL 3] - Contextual information to understand and work with electronic database of IPD - dataset specifications (metadata which describes the variable labels, variable descriptions, code lists and formats) - SAS programs [LEVEL 3] - SAS logs [LEVEL 3] - test outputs [LEVEL 2] - Original participant level records - filled out (completed) CRFs for all trial participants [LEVEL 3] - laboratory reports for all trial participants [LEVEL 3] - medical records and diagnostic reports for all trial participants obtained as part of trial procedures [LEVEL 3] - Documents related to clinical trials often created by trial sponsors - Marketing Assessments - Email correspondence - Meeting minutes - Records of the Data Monitoring Committee (also known as DSMB) e.g. adjudication committee	Peter Doshi	Johns Hopkins University School of Medicine

Line number	Comment and Changes proposed	Name	Affiliation
	- Records of the Data Monitoring Committee (also known as DSMB) e.g. adjudication committee	Peter Doshi	Johns Hopkins University School of Medicine