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

All statements in this advice are made under the assumption that CTAG1 rules for patient data confidentiality and anonymisation are applied and effective, and that CTAG5 legal rules and CTAG3 rules of engagement are strictly followed. As a consequence, there will be no more reference to the CTAG1, CTAG3 and CTAG5 rules in the rest of this advice.

1. The following definitions were agreed

1.1 This advice refers to all data recorded in a clinical trial (as part of documents and aggregated or patient level data) 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 European Medicines Agency (EMA) throughout the life-cycle of medicinal products.

It is recommended that the policy will be applied

- prospectively for future submissions to the Agency, which may include existing clinical trial data.
- to clinical trial data already available at the Agency (Level 1 and Level 2 data as defined in Section 3). The Agency will provide a time schedule for publication of these clinical trial data from submissions received before the policy comes into effect.

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 that provides further definition of the database metadata.

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.

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 un-necessary 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. Some views expressed that and data should be made available irrespectively.

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

- A cumulative indexed list of all clinical trials present in the submissions shall be provided (even if already available in the table of content of the submission dossier, as it may not be enough) 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 one of the following identifiers or a combination of these identifiers: the EudraCT number (but it would only cover trials conducted in Europe and it is not commonly referenced in journals and published articles), the US NIH clinicaltrials.gov registry number (commonly referred to in the literature), the ISRCTN registry number or a number provided by the applicant at the time of submission. Should there be no identifier a new one should be created. It is thought to be useful to be able to link back clinical trials to journal article information and to clinical registries (EudraCT, clinicaltrials.gov, ISRCTN registries).
- 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 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.
- Text documents containing data should be human readable and searchable by anyone requesting the data from the EMA.
- Patient-level data should be accompanied by associated documentation that allows a rapid and clear understanding of the data and how to process 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 additional information from either the marketing authorisation holder or 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 mark-up 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.
    - Readability with non-proprietary software should not be a mandatory requirement according members of Industry.
    - Inclusion of SAS programs is discouraged by members of Industry as they could represent substantial intellectual property of the marketing authorisation holder (MAH).
    - Of note, SAS versions more recent than Version 5 can only be opened by a SAS software.

3. What types of data are to be included and in what format

Assuming that data privacy protection has been ensured for all data made available publicly, 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 require extensive storage capacity.

- Members of Industry are of the opinion that the publication of these types of data may cause data privacy protection issues and therefore there should be no request to provide originals.

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

- Level 1: for each product, a full cumulative list of clinical trials, including a unique study identifier and basic information about the study (e.g., study title, interventions and indications); 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.
    - According to members of Academia, the list should also include information on the materials available for the study (e.g., full or abbreviated CSR).
- Level 2: 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]).
    - According to members of Academia, anonymised completed CRFs should be part of Level 2.
    - According to members of Industry, not all appendices should be included in Level 2; specifically, patient data listings should be excluded.
- Level 3: 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.
    - According to members of Industry, the following items should be removed from Level 3: 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.

The following table (Table 1) lists data elements, level of information, format in current submissions and whether they are routinely requested by the EMA. Advice was given by members of Industry that access to or publication of information will be in line with Level 1 and 3 definitions above, and any further request to data/information listed in Table 1 will require consideration on a reasoned case-b-case basis.

Table 1. Types of data

Type of data	Level	Routinely requested by EMA	Format
ICH E3 1-15 Core report	2	Yes	PDF
ICH E3 16.1 Study information	2	Yes	PDF

Type of data	Level	Routinely requested by EMA	Format
ICH E3 16.1.1 Protocol and protocol amendments	2	Yes	PDF
ICH E3 16.1.2 Sample case report form (unique pages only)	2	Yes	PDF
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	2	Yes	PDF
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	2	Yes	PDF
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	2	Yes	PDF
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	2	Yes	PDF
ICH E3 16.1.7 Randomisation scheme and codes (patient identification and treatment assigned	2	Yes	PDF
ICH E3 16.1.8 Audit certificates (if available)	2	Yes	PDF
ICH E3 16.1.9 Documentation of statistical methods	2	Yes	PDF
ICH E3 16.1.10 Documentation of inter-laboratory standardisation methods and quality assurance procedures if used	2	Yes	PDF
ICH E3 16.1.11 Publications based on the study	2	Yes	PDF
ICH E3 16.1.12 Important publications referenced in the report	2	Yes	PDF
ICH E3 16.2 Patient data listings (not all listings are routinely requested by EMA)	2	Yes	PDF
ICH E3 16.3 Case Report Forms	2	No	N/A
Investigator's brochure	2	No	N/A
Annotated Case Report Forms	3	No	N/A
Patient-level dataset (raw and derived)	3	No	N/A
Analysis datasets	3	No	N/A
Dataset specifications (metadata which describes the variable	3	No	N/A

Type of data	Level	Routinely requested by EMA	Format
labels, variable descriptions, code lists and formats)			
SAS programs	3	No	N/A
SAS logs	3	No	N/A
Test outputs	3	No	N/A
Completed CRFs for all trial participants	3	No	N/A
Laboratory reports for all trial participants	3	No	N/A
medical records and diagnostic reports for all trial participants obtained as part of trial procedures	3	No	N/A
Email correspondence	3	No	N/A
Meeting minutes	3	No	N/A
Records of Data Monitoring Committees	3	No	N/A

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. It should also be explored whether a database format is more suitable for Level 1 data.

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.

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

It was recognised that CDISC has defined useful formats: SDTM (Study Data Tabulation Model) for data tabulations, ADaM (Analysis Dataset Model) for analysis datasets, and define xml for metadata. The recommendation is for all these to be submitted to the Agency. 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 and FDA requirements in relation to these guides should be consistent.

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 the form of tables, e.g. in a comma-separated values (CSV) format; associated metadata should contain at least one table with all datasets, all

variables and their meanings, 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 not be 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. A transition period to provide the documents was recommended by group members.

- 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. EMA shall cooperate with the US FDA in the global development and alignment of formats, e.g. through CDISC and ICH. 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 a marketing authorisation application.

It is recommended that the International Organisation for Standardization (ISO), the European Committee for Standardization (CEN) and CDISC work together to define CSR 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

ISRCTN

The International Standard Randomised Controlled Trial Number (ISRCTN) is a simple numeric system for the identification of clinical trials worldwide.

http://www.nlm.nih.gov/bsd/policy/clin_trials.html

Annotated CRF

This is a blank case report form with annotations that document the location of the data with the corresponding names of the datasets and the names of those variables included in the submitted datasets.