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 4th February 2012 provided advice as follows:

General comments:

Term "format" - EFPIA would like to draw your attention to the fact that the meaning of the term is used variably throughout the document and would needs clarification (e.g. datasets formats, reference formats, metadata format, open file format).

1. The following definitions were agreed

1.1 This advice refers to all data recorded in a clinical trial (at a patient level or derived) that can be stored electronically and associated metadata (variable definition, terminology such as code lists or dictionaries) that are part of a submission for marketing authorisation to the Agency.

WHAT CLINICAL TRIALS ARE IN SCOPE?

Section 1.1 refers to data that are "part of a submission for marketing authorisation to the Agency". It is not clear to me what this means. Would this also cover data submitted after marketing authorisation, e.g. data submitted later in the life cycle of a drug? This could be data from studies submitted as part of PSURs or other updates on the evidence on a drug. I think all data submitted to the Agency should be made publicly available.

Proposed change: … that is submitted to the Agency.

Clarification of scope of application: EFPIA would appreciate a clarification that the new rules apply only to studies as included in submissions as of January 1, 2014 and beyond. While EFPIA agrees with EMA' s summary that, as a matter of principle, all clinical trials should be under the scope of the future policy, there is an urgent need to discuss the obligation of marketing authorisation holders when it comes to the submission/ reporting on clinical trials for which the MAH was not the sponsor, e.g. purely academic trials. As a matter of fact, in their submission MAH reference to publications but do not have the ownership on the underlying data for those studies which were performed without the MAH' s sponsorship and support. In those cases, MAH cannot be made responsible for the submission of data in the format set by the future policy.

Proposed change: "This advice refers to all data recorded in a clinical trial…..that a part of a submission for marketing authorization to the agency as of January 1, 2014 or beyond."

DEFINITION OF METADATA

There is currently no satisfying and agreed definition for "metadata", but vague ones (e.g., "data about data"). The definition suggested here is too restrictive, as you need much more than variable definition and codelists to have a proper metadata set (see for instance define.xml 2.0). Also terminology is one thing (a standard name for a given thing), code lists are other things (a set of choices for a question), although there is some overlap.

Proposed change: … and associated metadata (data properties such as dataset keys, variable definition, terminology, code lists).

Clarification needed whether the listing of metadata (variable definition, terminology such as code lists or dictionaries) is a complete list or represents examples. This should be at least specified in more details somewhere later.

Metadata should include also the context of the data interpretation, the rules chosen to code data, the hypothesis and the context of the study, the link between data and CRF and analysis.

Proposed change: associated metadata (any data useful to interpret the clinical data: variable definition, terminology such as code lists or dictionaries, the context of the study and the data, the purpose of the analysis, etc.)

1.2 Data formats refer to the organisation of information according to pre-set specifications that facilitate the storage, exchange and archive of clinical data. It includes both the type of electronic files and the content of the files, as well as associated metadata.

Data format should not refer to the content of the files, and should refer to organization of the data as noted in line 50.

Proposed change: Remove reference to content.

Need to exclude pdf formats.

Proposed change: that facilitate the storage, exchange, analysis and archive of clinical data.

This document refers to clinical data but is the intention to release computer programs as well. Where is the formatting of those to be considered?

The principles shall apply to clinical data submitted for regulatory submission throughout the life-cycle of medicinal products.

WHAT CLINICAL TRIALS ARE IN SCOPE?

Here the date/time point for the first release of data to a third party should be defined.

Again it is not clear to me, if "data submitted for regulatory submission" somehow restricts the data to be published. This should not be the case.

Proposed change: The principles shall apply to clinical data submitted to the Agency throughout the life-cycle of medicinal products.

The data and metadata concerned by this policy are stored and submitted electronically, but not necessarily sourced via electronic tools.

Even if not sourced via electronic tools, the data format must guarantee to link these documents to the data.

Proposed change: submitted electronically and guarantee the coherence of the data and documents even if not sourced via electronic tools.

The meaning of "sourced via electronic tools" is unclear. The requirement must be that the data itself is machine readable, but that requirement may not exist for the metadata.

Proposed change: The data and metadata concerned by this policy are stored and submitted electronically, but data must be machine readable but metadata may not need to be machine readable.

2. There is a need to define data formats

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.

There is an implication here that different formats may be requested for different purposes or customers. We strongly recommend to keep to the grandfather principle and not convert legacy data.

Proposed change: Formats as used by the company for the analysis may be different from study to study. Data should be made available in this format irrespective of the type of information to be made publicly available and the intended use of it ('grandfather principle').

As there are not universally agreed standards or formats, in order to avoid errors, a minimum set of rules should be defined, including:

- Indexed list of all trials present in the submissions shall be provided so the data of overall clinical program is tracked

The list of trials should allow to track the studies also in other systems that present data on a trial, e.g. in clinicaltrials.gov. Therefore, the list should include a unique trial identifier.

Proposed change: …clinical program is tracked and studies are identified by a unique study identifier.

Usually, companies are requested to provide the information already today as part of a submission.

Proposed change: "Indexed list of all trials present in the submissions shall be provided so the data of overall clinical program is tracked as long as not available in the table of context of the submission."

- Data shall be published in the format they have been submitted and evaluated

EMA may have received some approval file with all data in pdf files but the clinical studies have probably been analysed with adequate electronic files available by the firm. If the firm used proprietary software, it should change the format to a format for non-proprietary software.

Proposed change: Data shall be published in the format they have been submitted or evaluated by the marketing authorisation holder.

- Data should be readable and contain metadata to allow further analyses

Data should be readable - it is not clear by whom. Readability does not guarantee availability for analysis.

Proposed change: Data should be presented as a structured database.

Clarification needed through introduction of examples.

Proposed change: "Data should be readable and contain metadata to allow further analyses (e.g. SAS dataset format)."

Readable could apply to pdf. Could you propose another term saying that the data could be analysed with non-proprietary software such as openoffice.org spreadsheets (free Excel).

Proposed change: Data should be readable with a spreadsheet software such as openoffice.org one and contain metadata to allow further analyses.

I advise to add the notion that the metadata will provide the context to interpret correctly the data.

Proposed change: contain metadata that provide the context to interpret correctly the data and allow further analyses.

- Consistency with agreed terms throughout the life cycle of the medicinal products shall be maintained

Clarification needed what is meant by “agreed terms are”; is this the agreement on the data format that was originally agreed? Does this mean all studies for one product should be in the same format (which may be difficult for long lasting projects)? What is meant by “consistency” in this context?

- Formats at high level should be readable with electronic non-proprietary software

We should define a minimum standard which is realisable even in a small academic institution or a SME (e.g. scans of examination forms as PDF).

Clarification needed through introduction of examples.

Proposed change: "Formats at high level should be readable with electronic non-proprietary software. (e.g. reading SAS format)."

Formats at high level should be readable -not clear - data should be readable? What is meant by high level?

Proposed change: Remove or combine with line 104

The requirement to make the data readable with non-proprietary software contradicts the regulatory requirement that for regulatory relevant clinical studies only statistical software must be used which is validated and accepted.

3. What is to be included in data formats

Sponsors and institutions for statistical analyses are working with their own statistical software requiring sometimes standardized data formats, sometimes software specific data formats. If there is a mandatory data format the software cannot work with, they will have to buy new software or extra migration tools. To guarantee a future readability of older files and formats, especially the huge pile of data collected before this discussion, sponsors and institutions have to keep the old software in parallel. Reformatting existing data into a new format is costly, complex and a source for errors. The group should keep in mind, that this also applies to academic researcher and it seems doubtful, they can afford or pay for it.

Nevertheless, if this group intends to discuss an approach to a common usable study data file format, it should observe a very similar approach of the US FDA. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm248635.htm.

Proposed change: The sponsor should not be forced to migrate his own data collecting tools and statistical software environment to a new format as long as the electronic file used is a commonly used

format or can be migrated by the third party user into such one. I suggest to use a varying statement at this timepoint according to the paragraph, I copied from the current FDA draft guidance on eStudy data format:

A file format standard specifies a particular way that information is encoded in a computer file. Specifications for a format permit the file to be written according to a standard, opened for use or alteration, and written back to a storage medium for later access. Some file formats in widespread use are proprietary, others are open source. Examples of file format standards supported at FDA include Adobe Acrobat Portable Document (.pdf), SAS Transport File format (.xpt), text files (.txt), and Extensible Markup Language (.xml). The use of a file format standard for study data exchange supports technical interoperability, but by itself is often insufficient for semantic interoperability.

There is an absolute need that formats agreed contribute to ensure privacy protection. Certain information such as CT scans, MRI and other imaging, interviews shall not be included in the formats as they carry too many identifiers. Without appropriate guarantees public disclosure of clinical data might have a negative impact on recruitment.

IN FAVOUR OF INCLUDING CT scans, etc.

CT scans, MRI, interviews etc. should not be excluded per se - this will jeopardize the whole CT Data Transparency idea. Many CT scans only show a small body region and if the name of the patient is replaced by the study ID than there will be no possibility to identify him the person.

Proposed change: … and other imaging, interviews should be carefully checked so that they contain no data which might be used to identify the patient.

Patients' genetic/genomic data should not be included either.

Proposed change: Add patients' genetic/genomic data to the exclusion list.

Whilst I agree that CT Scans, MRI and other information may compromise the data privacy, I believe the results from these should be included and would not compromise the data privacy.

For imaging, interviews...: their analysis leads to a full set of data written in a specific CRF (volume, number, position of lesions, characteristics...). All rules that apply to the "clinical/biological" CRF should apply to the imaging, interview CRF.

NOT IN FAVOUR

As discussed at the meeting at EMA in November 2012, requesting absolute privacy protection might make publication of any data impossible. Standards set by the European data protection officer should be considered sufficient.

Data Privacy should explicitly be mentioned to apply to genetic data. In addition, respecting data privacy goes beyond CT Scans, MRI or other imaging. There is a need for a reference to the EU data protection Directive 95/46/EC and to anonymization, in particular through de-identification, removal of free text, date of birth anonymization, obfuscation of subject study dates.

Proposed change: "There is an absolute need that formats contribute to ensure data privacy protection through anonymization (reference to Directive 95/46/EC). Obviously, certain information such as CT scans, MRI and other imaging, interviews and genetic data shall not be included in the formats as they carry too many identifiers."

Three levels of clinical data and corresponding formats shall be included

Items should be listed, but it may be necessary to list E3 items.

Proposed change: Full CSR including protocol, amendments, dated SAP, CRFs, individual level data, certificates of analysis, list of IRB and IC, supplementary tables, informed consent forms, list of investigators, list of contributors to CSR.

Three levels of study information, data and corresponding formats shall be included.

Level 1: full list of trials of a given drug including a unique study identifier for each study; these lists should be fully searchable; the lists could be connected to the EPARs.

Level 2: for each study: full clinical study report (CSR) according to ICH E3 including all appendices (this format according to ICH E3 among other things includes a full protocol with all amendments, a full statistical analysis plan and full summary tables and test outputs). A report according to ICH E3 also includes patient data listings. Measures needed to protect privacy to be discussed.

Level 3: for each study: data sets (including individual patient data) and results used for the evaluation of the drug (including meta-data required to use the data set, like an annotated CRF, variable definitions, derived values etc.); including any test outputs.

- Full clinical study reports: acceptable in PDF format for all approved medicinal products.

Need to also make full study protocols with all amendments dated available with CSRs, otherwise the reporting of appropriate outcome measures, statistical analysis plans, populations to be evaluated etc. in CSRs cannot be assured.

Proposed change: Full clinical study reports plus complete study protocols (including all amendments)…

Why to narrow that on "all approved medicinal products" - this should also fit (if applicable) for other issues (e.g. MP under investigation; MP having not yet an approval).

Full clinical study reports must not include patient level data unless anonymized.

Proposed change: “Full clinical study reports (excluding individual patients’ level data).”

- Datasets and results used for the evaluation linked to the relevant protocols; full statistical analysis plan, details on methods and metadata are to be made always available to allow a meaningful re-assessment.

Datasets and ..... requires clarification.

Proposed change: Patient level datasets and .....

Clarification: Results are already submitted with the study reports.

Proposed change: “individual patient data sets used for the evaluation linked to the relevant protocols; ….”

Full data should include all the data obtained on the CRF even if the data have not been analysed for the study report (as far as confidentiality is not engaged). For any specific subgroup of patients defined for analysis or as a consequence of an analysis (patients considered for per protocol analysis, responders vs non responders, patient with a specific characteristic or outcome) this should be clearly

indicated in the dataset on a patient basis. For example a column should indicate if yes or no the patient should be considered in the group.

Proposed change: Datasets and results used for the evaluation linked to the relevant protocols; full data included in the CRF (except confidential information), full statistical analysis plan, details on methods and metadata are to be made always available to allow a meaningful re-assessment. For any subgroup of patients defined in the full clinical study report, dataset should include information on a patient basis on whether or not the patient is belonging to the group.

- Individual data such as CRF in PDF format are neither useful (as they will require substantial manpower for reloading in another usable format) nor appropriate as may contain subjects identifiers breaching privacy protection. Data from the annotated CRF are to be included.

SAS files should be made available for IPD.

I disagree with the first sentence! PDF scans of printed out CRFs are by far ideal for reassessment of data but might be the minimal standard which is realisable even in a small academic institution or a SME. We should not forget that data transparency is the second step after generating the data in the hospital (first step) - and we should not set up unnecessary burden for financially weak "small sausage" holders (which cannot spend the money for a steak).

Clarification: annotated CRFs never contain patient data.

Proposed change: delete: "data from"; change into: "The clinical trial data should be accompanied by an annotated CRF."

Some old files probably contain datasets in pdf format. If yes and if the requester wishes, EMA should ask the marketing authorization holder to provide dataset in a format that can be used in spreadsheets.

Proposed change: Individual data such as CRF in PDF format are neither useful (as they will require substantial manpower for reloading in another usable format) nor appropriate as may contain subjects identifiers breaching privacy protection. Data from the annotated CRF are to be included in the format. If the MA file contains data in pdf format, EMA should ask the MAH to provide the data in an adequate format readable in a non-proprietary spreadsheet format.

I strongly agree that the annotated CRF should be submitted. If CDISC SDTM Data is to be used as a pre-requisite or guide then the SDTM annotated CRF would also be very useful for the reviewer. SDTMs and ADaMs should both be submitted. Question is would the Raw CRF data be useful to EMA or is SDTM sufficient?

More detailed discussion is needed on what additional elements shall be provided along with the datasets.

There may be circumstances that would justify a different assessment of the confidentiality of the clinical trial data (this is being discussed in CTAG5 about legal aspects) and in such cases the level of clinical data and corresponding formats may need to be adjusted accordingly.

It would be good if such “additional elements” could also be harmonized, especially with FDA and eSUB requirements.

Proposed change: "More detailed discussion is needed on what additional elements shall be provided along with the data. Harmonization with other agencies should not only be achieved with regard to data structures and formats but also with respect to such additional requirements."

A general comment about unstructured data. Unstructured data have to be managed in order to enhance their usability. The formats should include the links between structured data and unstructured data. This will help at every step, from analysis to review.

Proposed change: All three levels of clinical data should be tightly linked in order to guarantee their readability and their usefulness.

4. Formats recommended

For the clinical study reports, the full documentation shall be made available according to the ICH E3 format.

Clinical study reports should include appendices; possible measures with regard to privacy protection with regards to patient data listings to be discussed.

Proposed change: For the clinical study reports, the full documentation shall be made available according to the ICH E3 guideline, including appendices.

CSRs must be identical to the original document, signed and dated by the sponsor.

Proposed change: For the clinical study reports, the full documentation in its original version signed and dated by the sponsor….

Old clinical study reports may not fully comply with the current ICH E3 format. In these cases it should be acceptable to provide the clinical study report in the original format in which it was written.

Proposed change: For the clinical study reports, the full documentation shall be made available according to the ICH E3 format, or the original format in which the report was written.

To avoid delays any format shall be acceptable for products already authorised. The data shall be published in the format they are available at present then the format could move progressively to CDISC. However, CDISC provides a frame but for the data itself, there are no agreed standards: those shall be developed gradually applying the grandfather principle.

In the meeting on 4th Feb 2013 EMA confirmed that the policy would be applied prospectively for new medicinal products after the implementation date therefore products already authorised should be out of scope. The wording should allow for release of old clinical data included in new MAAs.

Proposed change: To avoid delays any format should be acceptable for active substances contained in authorised medicinal products.

Datasets may already be available in CDISC formats.

Proposed change: …….published in the format they are available at present, including CDISC, then the format could move progressively to CDISC as recommended.

CDISC does provide standards for the data to be collected.

Proposed change: Remove the sentence 'However, CDISC provides a frame but for the data itself…….

Proposed change: "To avoid delays any formats should be acceptable for those studies which have already been started at the point of entry into force of the new policy. The data shall be published …"

CDISC could be a useful format for datasets, but for metadata other formats might be more useful.

WHAT CDISC STANDARDS?

CDISC is not a format, but an organization. Please clarify what you are talking about here: ODM? SDTM? ADaM?

CDISC STANDARDS FOR METADATA

It seems odd to talk about "other formats" for metadata. CDISC has developed not only good and widely used standards for data, but also good and compatible standards for metadata, in the form of define.xml. As CDISC standards are already widely used in clinical research, it would be highly desirable to make use of them to the fullest extent possible and not to reinvent the wheel.

Proposed change: Replace line with "CDISC have defined useful formats for both data, in the form of SDTM and ADaM standards, and metadata, in the form of define.xml. Use of these standards is strongly encouraged."

CDISC standards for metadata are well defined and work well with CDISC formatted data.

Proposed change: CDISC could be a useful format for datasets, but for metadata other formats may also be considered.

OTHER STANDARDS FOR METADATA

EMA should consider minimal requirements for metadata.

Proposed change: "CDISC could be a useful format for datasets. EMA should define minimal requirements and standards for metadata."

Whilst CDISC formats provide a good guide to data formats, there remains much ambiguity over the Implementation Guides of CDISC with many Pharmaceuticals adopting their own interpretation. This would be a good opportunity to resolve this ambiguity and create clear and concise definitions.

I suggest adding a sentence to indicate the direction taken: multiple standards. Therefore, there is a need of a standard to link the different standards. CDISC and HL7 propose BRIDG.

Proposed change: Therefore, it seems that there is a need of a set of standards, for each type of data or exploitation.

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 to ensure consistency and versioning control.

An explanation is needed what is meant with “in a collaborative way” and who would be included.

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.

This seems to mean "the CDISC is evolving SDTM way too fast, so the EMA should not follow this pace". If it is what is meant, please say it explicitly.

Whatever the format chosen, dataset formats in the long term are to be compatible.

Please define "compatible": with what? This word has no meaning alone.

Clarification needed on what compatibility means in this context.

Proposed change: "Whatever the format chosen, dataset formats in the long term are to be compatible in the standard format being used (e.g. like CDISC)."

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 trials are included.

EFPIA agrees to the need to accept long term studies with different formats attached when studies were finished at completely different time points.

Old non-formatted data in older studies may be an issue and could prove difficult to re-format to a newer version if required.

5. Who should adhere to the agreed formats

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

Local studies outside Europe (e.g. local registration studies in Korea, China, Russia, Ethiopia etc.), will only be expected to adhere to the agreed data formats, if they are part of a submission to the EMA. In any case, international harmonisation of data formats is needed before submission of trials in the future EU data formats can be required.

Proposed change: "The formats agreed are to be adhered to by all stakeholders and also for trials run outside Europe if they become part of a submission to EMA."

For trials owned in different measure by different partners (e.g. public-private partnerships), the above points should be taken into account from the beginning of the clinical studies.

The situation regarding observational studies conducted by third parties requires further consideration and discussion. There are strict rules in place regarding industry use of third-party data. In these cases the Marketing Authorisation Holder is not permitted to share the data. Data from observational studies should be exempt from disclosure.

Additional comment on partnership programs.

Proposed change: "This concerns a potential inconsistency between data submitted to the Agency for which the MAH takes accountability. In addition, it concerns publications which could fall under the remit of a public-private partner and which could use for example a different data cut. A clarification is needed on how agreements of public-private partners on secondary publications can be maintained when studies are being made available at the time of approval in agreement with the MAH."

The scope of phase 4 clinical trials should be clarified as the MAH does not always have access to these data. We assume this applies only to studies conducted by the MAH which would be submitted to the MAA.

6. Timelines for format implementation

The CTAG2 recommended the policy to be implemented from January 2014.

The recommendation to implement the policy from January 2014 includes publication of data already available at the Agency before January 2014 to be pro-actively published starting January 2014. Clinical study reports of all approved drugs available at the agency from submissions before January 2014 should be published by the Agency. These CSRs are required to assess drugs in current use beyond the assessment provided by the regulatory agencies for marketing authorisation. Examples for additional assessments which could be informed by these CSRs are questions of reimbursement or indirect comparisons required for comparative effectiveness research.

Proposed change: The CTAG2 recommended the policy to be implemented from January 2014. It is furthermore recommended to pro-actively publish also those CSRs which are available at the Agency from submissions before January 2014. Publication of CSRs submitted before January 2014 should also start in January 2014.

Clarification on applicability of the new policy.

Proposed change: "The CTAG2 recommended the policy to be implemented in all submissions from January 2014 onwards."

This part supposed that we act on this dates. During TC I understood it was a proposed estimation.
Proposed change: Replace recommended by evaluate or suggest.

- Clinical data for products already approved to be published in the format available at the time of submission.

Should clarify that all CSRs previously submitted to EMA for approved drugs will be made available.
Proposed change: Clinical data regarding all trials submitted to EMA for products already approved…

Clarification of scope and timing for the future policy.

Proposed change: "Anonymized clinical data from studies already started before implementation of this policy (and especially for those already analysed) should be published in the format available at the time of submission."

Clinical data for products already approved can be published in the format available at the time of submission. Clinical study reports of all approved drugs available at the agency from submissions before January 2014 should be published by the Agency. These CSRs are required to assess drugs in current use beyond the assessment provided by the regulatory agencies for marketing authorisation. Examples for additional assessments are questions of reimbursement or indirect comparisons required for comparative effectiveness research.

- Data for new marketing authorisation submissions to be made available in an open file format.

Please clarify again what is meant with "open file format", by adding an example like "e.g. SAS transport files".

- Pro-active adoption of standard formats: as this has to be mandatory for the sake of fairness and clarity for all stakeholders, it was advised starting gradually to acquire experience and then mandate formats after 2-3 years of trial period.

While it seems reasonable to gain experience with formats of data sets and individual patient data, there is no need 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 can be mandatory starting January 2014.

Proposed change: … after 2-3 years of trial period. Since the format of clinical study reports is established since 1996, the CSR format (ICH E3) becomes mandatory in January 2014.

More clarification needed what is meant with "…2-3 year of trial period".

Proposed change: "…, it was advised starting gradually to acquire experience and then mandate formats after 2-3 years for all new studies."

7. International harmonisation across regulatory agencies

Agree that harmonisation is required but also should implement what will be widely used in future to further standardise the process and prevent any re-formatting.

EFPIA considers global alignment and harmonization are critical steps in the future process.

EMA is leading in terms of policy but global consultation of formats is recommended.

The level for global consultation should be ICH (VICH resp.)

Proposed change: … but global consultation of formats at the ICH (VICH resp.) level is recommended.

I would like to propose a "probationary time period" during which the consequences of the strategy agreed upon could be re-assessed and in case some not expected consequences will turn out to be inadequate, adaptations could be performed before such complex new rules will be implemented.

Given the requirements from FDA on eSUBs, a recommendation on global consultation of formats is insufficient. Multiple formats will result in duplicative work and unsustainable burden on industry.
Proposed change: "EMA is leading in terms of policy but global consultation of formats is essential to ensure that only one format is required to be produced by industry for regulators worldwide."

The international harmonisation is a critical point, as formatting data is costly and, moreover, maintaining integrity of data in several formats is not a good practice that can lead to errors. EMA with FDA, SFDA and Japan PMDA, and any other public agency, should harmonise their recommendations on data formats and metadata requirements.

Global alignment for both the initial agreed formats and for the updates are necessary.

Under e-CTD, PDF, XML and other standards are allowed in MAA.

ISO, CEN and CDISC to define CSRs harmonised standards.

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
7	Comment: The sentence (as part of documents and data aggregated or at patient level). **Proposed change (if any)**: Should be (as part of documents and aggregated data or patient level data).	Mary Sinnathamby	Parkinson's UK
7	Comment: There are a lot of acrynoms used in section 4 which not everyone may be familiar with such as SDTM and ADaM. **Proposed change (if any)**: Include a glossary section (particularly as the document is going to be opened for public consultantion) or add the full terms when used for the first time.	Mary Sinnathamby	Parkinson's UK
11	Comment: I agree with the definition stating that the advice refers to all data submitted throughout the life-cycle of a drug.	Beate Wieseler	IQWiG

Line number	Comment and Changes proposed	Name	Affiliation
11	Comment: The following sentence is not clear to me: "The policy will be applied prospectively for future submissions to the Agency; it may include old clinical trial data." The scope of the advice should be defined as follows: the advice should apply to all clinical trial data submitted to the Agency (i.e. clinical trial data submitted before January 2014 and clinical trial data submitted in and after January 2014). For clinical trial data submitted before January 2014 only study lists and clinical study reports (CSRs) can be published because data sets are not available at EMA for these trials. For clinical trials submitted after January 2014 EMA should routinely require submission of data sets (including metadata) for publication. Full trial information should routinely be proactively published after a decision by the Agency (independent of whether the decision is negative or positive). **Proposed change (if any)**: The policy will be applied prospectively for future submissions to the Agency (which may include old clinical trial data). In addition, the policy will apply to clinical data already available at the agency (Level 1 and Level 2 data as defined below). The Agency will provide a time schedule for publication of these clinical trial data from submissions received before the policy comes into effect.	Beate Wieseler	IQWiG
11	Comment: Clarification is needed with respect to the application for future submissions related to indications already approved for use prior to January 2014. **Proposed change (if any)**: " The policy will be applied prospectively for future submissions to the Agency, including those related to already approved indications, and thus may include old clinical trial data."	Sabine Atzor, Hans Ulrich Burger	EFPIA

Line number	Comment and Changes proposed	Name	Affiliation
12	Comment: Should be more precise. Will it be reformated data, or an electronic scanned version of documents already provided to EMA? **Proposed change (if any)**: old clinical trial data will be considered separetly, in a specific workshop dedicated to legacy data.	Patrick Lamplé	Institut de Recherches Internationales Servier
12	Comment: The statement: "it may include old clinical trial data." is vague, as 'old' is not defined. **Proposed change (if any)**: Clinical trial data will be included for drugs approved since 1995.	Barbara Mintzes	Health Action International
20	Comment: I agree that legal requirements for for patient data confidentiality and anonymisation are to be followed. However, if these requirments affect the content of CTAG2 advise (e.g. by making delivery of an agreed data format impossible), CTAG2 should be informed about this to be able to provide advise on alternative solutions with regard to data formats. **Proposed change (if any)**: 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. If any of these requirements affect the provision of data formats described in this advise, CTAG2 will be informed to be able to provide advise on alternative solutions with regards to data formats.	Beate Wieseler	IQWiG
20	Comment: We are not officialy informed of the ongoing discussion of CTAG1 and CTAG5. As far as we know, no definitive documents were published. Please provide the documents and rules that you refer to and that we must follow and apply. **Proposed change (if any)**: to be removed.	Patrick Lamplé	Institut de Recherches Internationales Servier

Line number	Comment and Changes proposed	Name	Affiliation
20	Comment: There are major concerns that data confidentiality and anonymisation can be applied and be effective when data is being made available to everyone in an uncontrolled process. We understand that this is not the right forum to discuss details but the topic has to be raised as it is significant and the text should be changed to an "assumption". **Proposed change (if any)**: "All statements in this advice are made under the assumption…".	Sabine Atzor, Hans Ulrich Burger	EFPIA
20-23	Comment: The results of CTAG 2 should be read in conjunction with all other CTAGs, in particular CTAG 1, 3 and 5. **Proposed change (if any)**: add: "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 CTAG 5 legal rules and CTAG 3 Rules of engagement are strictly followed."	Sabine Atzor, Hans Ulrich Burger	EFPIA
25	Comment: Please be more precise for the words "delays" and "unneccessary". Any standards, including format, induce delays and effort to format and exploit the data. **Proposed change (if any)**: to be removed.	Patrick Lamplé	Institut de Recherches Internationales Servier
28	Comment: Please precise what is proposed by "made available irrespectively".	Patrick Lamplé	Institut de Recherches Internationales Servier

Line number	Comment and Changes proposed	Name	Affiliation
29	Comment: True, but CDISC formats are international and majoritary (if not the only international format available) and used in many countries. EMA should clearly specify if it intends to participate in CDISC initiative. It could be part of the recommandation of the CTAG2. **Proposed change (if any)**: To be removed and put a link to the end of the document where this topic is also discussed.	Patrick Lamplé	Institut de Recherches Internationales Servier
30	Comment: An indexed list of all trials including a minimum description of the trial (unique identifier, study title, interventions, indication) should be made available in all cases (Level 1 data described below). It would be insufficient to only have information on the study programm from a table of content of the submission dossier, because this would be impossible to properly search and handle. **Proposed change (if any)**: An indexed list of all clinical trials present in the total of all submissions including a minimum description of the trial (unique identifier, study title, interventions, indication; Level 1 data described below) shall be provided so the data of the overall clinical program is tracked.	Beate Wieseler	IQWiG

Line number	Comment and Changes proposed	Name	Affiliation
30	Comment: Even if the list is present in the submission dossier, the list should be available anytime, with the same format and details. **Proposed change (if any)**: "if not already" should be replaced by "even if already".	Patrick Lamplé	Institut de Recherches Internationales Servier
31	Comment: (if not already available in the table of contents of the submission dossier) It would be better to list trials per drug in a consistent manner, as table of contents listings may be incomplete or organized in a variety of ways. If trials are being made available similarly for all products, a consistent, centrally available list would not create much more work and would be very useful for users. This provision is also specified already in lines 68, 69 and 70. **Proposed change (if any)**: Please remove this clause.	Barbara Mintzes	Health Action International
33	Comment: This identifiers should be present, and another unique identifier should be created, specific to EudraCT results submission process, in order to have a common and unique identifier. **Proposed change (if any)**: "This identifier could be either:" should be "This identifier will be unique and combine with either of this other identifiers:"	Patrick Lamplé	Institut de Recherches Internationales Servier
36	Comment: the identifyer should not be an internal number provided by the applicant. it would be better, for the sake of cross-referencing, to use a clinical trial identifier from a pucly and recognised database, like the examples mentioned supra. **Proposed change (if any)**: remove this part of the sentence.	Bertrand Le Bourgeois	Medidata Solutions

Line number	Comment and Changes proposed	Name	Affiliation
38	Comment: I agree that no conversions should be required for publication of data. However, it needs to be ensured that the submission made to the Ageny meet the minimum requirements defined in this advise (concerning content and format). **Proposed change (if any)**: 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). The Agency will specify the minimum requirements defined for publication of clinical trial data in this advise as submission requirements for future submissions.	Beate Wieseler	IQWiG
38	Comment: How this will be managed compare to line 12 about old clinical trials data?	Patrick Lamplé	Institut de Recherches Internationales Servier
40	Comment: Clarification that modification of the format may be possible to ensure redaction. **Proposed change (if any)**: Add: “Appropriate, quality-controlled modification of the file format may be performed to ensure redaction of commercially confidential information and personal data.”	Sabine Atzor, Hans Ulrich Burger	EFPIA
43	Comment: Should be more precise. SDTM is not a human readable format. It requires knowledge in CDISC SDTM format and an access to the SDTM referential used. Search functionalities mostly depends on the tool, not directly the documents. **Proposed change (if any)**: to be removed.	Patrick Lamplé	Institut de Recherches Internationales Servier
45	Comment: My feeling is that this sentence is a bit redundant. Excel is covered by the "open source, non-proprietary software (but not necessary free)"(line 54) and can be coverted in a CSV format (if needed).	Steven Deleu	EORTC

Line number	Comment and Changes proposed	Name	Affiliation
45	Comment: excel is a proprietary software. as said in line 54, the text should stay with open source products. **Proposed change (if any)**: ...that patient-level data could be exported to spreadsheet softwares.	Bertrand Le Bourgeois	Medidata Solutions
45	Comment: I am not sure it would be advisable to analyse a clinical trial data base in Excel.	Beate Wieseler	IQWiG
45	Comment: Too restrictive. **Proposed change (if any)**: Patient data level should be available in tabular format.	Patrick Lamplé	Institut de Recherches Internationales Servier
45	Comment: Reference to a single request for performance of patient level data analysis in excel is a stand-alone and not connected to subsequent statements. In addition, as to our knowledge, excel is not aimed at producing valid statistical results. This means an analysis conducted in excel could lead to different results than one conducted with validated statistical tools such as SAS. **Proposed change (if any)**: delete bullet point.	Sabine Atzor, Hans Ulrich Burger	EFPIA
46	Comment: "Quiclky grasping the data and processing it" is too interpretable. Should be removed. **Proposed change (if any)**: to be removed.	Patrick Lamplé	Institut de Recherches Internationales Servier
46	Comment: Individual patient level data are complex and there is no way that allows "quickly grasping the data and processing it", independently of associated data documentation. We should not raise wrong expectations. Good documentation is needed so that data can be understood and processed. **Proposed change (if any)**: “Patient level data should be accompanied by associated documentation that allows its understanding and processing …(and add at the end line 53:). Necessary steps to ensure data confidentiality would need to be defined by EMA.”	Sabine Atzor, Hans Ulrich Burger	EFPIA

Line number	Comment and Changes proposed	Name	Affiliation
48	Comment: Aspect of machine-readability needs further clarification. **Proposed change (if any)**: "This documentation, which includes metadata (="structured data about data"), should ideally be machine-readable, i.e. CDISC Define.xml or pdf."	Sabine Atzor, Hans Ulrich Burger	EFPIA
52	Comment: Sentence correction: "for" can be deleted and double negation. Without…either…or.. **Proposed change (if any)**: "without needing additional information from either the marketing authorisation holder or the EMA.	Steven Deleu	EORTC
56	Comment: As to our knowledge SAS versions more recent to Version 5 can only be opened by SAS software. Furthermore, SAS programs should not fall under transparency since they could represent substantial intellectual property of the MAH, especially when generic analysis systems are used for generating large parts of study results. (also see comments to line 76 and 77). **Proposed change (if any)**: "Formats…: that includes, but is not limited to….SAS transport file format up to version 5 (xpt) for datasets (delete: programs (as opposed to SAS format which is proprietary)) …."	Sabine Atzor, Hans Ulrich Burger	EFPIA
58	Comment: During the discussion it was proposed to have an international effort to have common requirements, and potentially managed by ICH. This harmonisation will therefore ensure common formats and standards. **Proposed change (if any)**: refer to the end of the document about international harmonisation.	Patrick Lamplé	Institut de Recherches Internationales Servier
62	Comment: cCertain **Proposed change (if any)**: certain	Steven Deleu	EORTC

Line number	Comment and Changes proposed	Name	Affiliation
62	Comment: Clarification that only summary data from CT scans, MRI and other should be in the scope for privacy reasons (as mentioned by the EFPIA delegation during the tcon on March 07). **Proposed change (if any)**: " Assuming that data privacy protection has been ensured for all data made available publicly, certain information such as CT scans, MRI and other imaging, interviews, genetic/ genomic data can bring useful information and only summarised data derived from such information should be in the scope of discussion for data formats."	Sabine Atzor, Hans Ulrich Burger	EFPIA
65	Comment: Adapt the line on storage size needed. **Proposed change (if any)**: "might cause serious informatics problems." should be change into "requires extensive storage capacities".	Patrick Lamplé	Institut de Recherches Internationales Servier
67	Comment: These three levels apply to post-market surveillance studies as well as clinical trials submitted prior to marketing. **Proposed change (if any)**: specify that this applies to trials throughout the product life cycle.	Barbara Mintzes	Health Action International
67	Comment: Unclear what these level apply to. **Proposed change (if any)**: These levels apply only to trials provided to EMA.	Tom Jefferson	Attentiallebufale, Italy
68	Comment: the minimum information for the list of trials should be specified. **Proposed change (if any)**: Level 1: 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 and could be connected to the European Public Assessment Reports. This is separate to information stored in the EUdraCT database.	Beate Wieseler	IQWiG

Line number	Comment and Changes proposed	Name	Affiliation
70	Comment: Unclear purpose of first level. **Proposed change (if any)**: The first level would be a searchable index and it will include all studies submitted to EMA in the lifecycle of medicines (at the time of MAA as well as after initial MA is granted).	Tom Jefferson	Attentiallebufale, Italy
71	Comment: Level 2 and Level 3 data should be tighly linked in order to prevent reinterpretation of level 3 data outside the context and the definition provided in level 3 metadata and level 2 description in the report. **Proposed change (if any)**: The three levels of data must be consider together. No reinterpretation of level 3 data can be done without specifying the context of the data, which is at least present in level 2 data.	Patrick Lamplé	Institut de Recherches Internationales Servier
71	Comment: CSR per se do not fulfil data privacy - this is a crucial point. It has to be defined what needs to be taken of a CSR to ensure data confidentiality like deducted patient narratives. In addition, the CSR may also include commercially confidential information which needs to be redacted. **Proposed change (if any)**: "Level 2: for each study, full clinical study report (CSR) according to ICH E3, including all appendices (study information, patient data listings and blank case report forms (CRF)) and which is deducted to ensure patient confidentiality (such as removing patient narrative) and to protect commercially confidential information in line with specifications provided by EMA."	Sabine Atzor, Hans Ulrich Burger	EFPIA
72	Comment: Very unclear whether listings are the same thing as IPD. **Proposed change (if any)**: Clarify difference (if any) between listings and IPD.	Tom Jefferson	Attentiallebufale, Italy
72	Comment: Item missing: Investigators' brochure. **Proposed change (if any)**: Investigators' brochure.	Tom Jefferson	Attentiallebufale, Italy

Line number	Comment and Changes proposed	Name	Affiliation
73	Comment: Unclear whether CRFs are specimen forms (i.e. empty) or not. **Proposed change (if any)**: Clarify.	Tom Jefferson	Attentiallebufale, Italy
76	Comment: Clarification of listing to avoid duplications; in addition SAS readable programs contain stakeholder/ MAH macros and their submission would require heavy preparatory work. This information should only be made available on justified grounds (such as analysis of efficacy data) and on specific request. Furthermore, SAS programs should not fall under transparency since they could represent substantial intellectual property of the MAH, especially when generic analysis systems are used for generating large parts of study results. **Proposed change (if any)**: "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, dataset define files). (delete reference to variable definitions, data derivation specifications, test outputs, SAS logs and SAS programs)".	Sabine Atzor, Hans Ulrich Burger	EFPIA
77	Comment: We do not believe test outputs should be part of the metadata as these are not part of a submission and only kept as part of program validation, which, in terms of using general tools, may not even be study specific. The same is true for SAS logs. SAS programs should not fall under transparency as explained for line 76. **Proposed change (if any)**: Delete "test outputs, SAS logs and SAS programs".	Sabine Atzor, Hans Ulrich Burger	EFPIA

Line number	Comment and Changes proposed	Name	Affiliation
79	Comment: If modifications of the data formats in Levels 1 to 3 are required based on confidentiality or legal requirements, this should be discussed in CTAG2. CTAG2 should be able to provide alternative solutions based on the legal requirements. Otherwise, the outcome of this advise could be changed substantially without input from the group, which puts the whole discussion into question. **Proposed change (if any)**: If elements included in the three levels of data listed above may need to be modified in special circumstances driven by confidentiality or legal aspects, this will be discussed in CTAG2.	Beate Wieseler	IQWiG
83	Comment: For future submission, data should comply with the minimum requirements specified in the advise. **Proposed change (if any)**: The data shall be published in the format they are available at present. For future submissions, data should be avilable in the formats described in the advise.	Beate Wieseler	IQWiG
84	Comment: The list of studies (Level 1) would be a living documents that needs to be update with every submission that includes a new study. **Proposed change (if any)**: In terms of the different types fo data described in the previous section, Level 1 data need to be updated with every submission including a new study and should be searchable.	Beate Wieseler	IQWiG
92	Comment: CDISC have **Proposed change (if any)**: CDISC has	Steven Deleu	EORTC
94	Comment: "but not ODM, which is a transport format for data management" is not accurate. ODM is a format for transporting clinical trial data. The sentence here is a bit confusing. Can ODM files be submitted/allowed?	Steven Deleu	EORTC

Line number	Comment and Changes proposed	Name	Affiliation
94	Comment: Clarification needed that data organised is in SDTM and ADaM within a SAS Transport v5 file as opposed to SDTM and ADaM with an ODM file. SAS/ODM specify file formats while SDTM/ADaM specify organisation of data. **Proposed change (if any)**: "The recommendation is for all these to be submitted to the Agency." (delete: "but not ODM, which is a transport format for data management.")	Sabine Atzor, Hans Ulrich Burger	EFPIA
99	Comment: Provide clarification on what "rectangular" means. **Proposed change (if any)**: No further proposal at this stage.	Sabine Atzor, Hans Ulrich Burger	EFPIA
101	Comment: also possibly can be deleted **Proposed change (if any)**: …their meanings, associated code lists,….	Steven Deleu	EORTC
107	Comment: "….re-formatting of old data should be not….". **Proposed change (if any)**: "….re-formatting of old data should not be….".	Mary Sinnathamby	Parkinson's UK
114-117	Comment: Flexibility is needed for stakeholders in the choice of the CDISC version and a corresponding clarification in the text. **Proposed change (if any)**: No proposal at this stage.	Sabine Atzor, Hans Ulrich Burger	EFPIA

Line number	Comment and Changes proposed	Name	Affiliation
128	Comment: Please clarify the target number of language. One should be enough, considering extracost for global and acurate translation (verbatim, symptoms, etc.)	Patrick Lamplé	Institut de Recherches Internationales Servier
133	Comment: The requirement to provide data according to the advise should cover all trials for which a marketing authorisation holder is the sponsor or in which the MAH is finanically engaged. If a marketing authorisation holder provides funding for a study, the data should be available as described in the advise. Such a requirements would foster appropriate contracts (e.g. in public - private partnerships).	Beate Wieseler	IQWiG
133	Comment: For ongoing studies, providing full data transparency may require a change in the contract with an academic partner. Therefore, clarification may be needed that the selection of standards along with further conditions should be agreed between the multiple partners on a case-by-case basis, e.g. through a contract. **Proposed change (if any)**: 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 data to the extent and according to the standard as detailed in the contract."	Sabine Atzor, Hans Ulrich Burger	EFPIA
133	Comment: What provisions will be made for studies carried out during the post-market period that must be reported in PSURs but have not necessarily been conducted by the manufacturer?	Barbara Mintzes	Health Action International
138	Comment: "can be mandatory from the implementation of the policy." At least a transition period should be defined, in order to be ready to provide all required data. **Proposed change (if any)**: "can be mandatory from the implementation of the policy, with a transition period to provide the documents".	Patrick Lamplé	Institut de Recherches Internationales Servier

Line number	Comment and Changes proposed	Name	Affiliation
145	Comment: Industry expects a strong commitment of EMA to cooperate with the US FDA in the global development and alignment of formats through ICH and CDISC. **Proposed change (if any)**: "The EMA is leading in terms of policy but global alignment and harmonisation are critical steps in the future process. (add sentence:) EMA will cooperate with the US FDA in the global development and alignment of formats through CDISC and ICH. A global consultation of formats is recommended…."	Sabine Atzor, Hans Ulrich Burger	EFPIA
N/A	Comment: Bracket missing **Proposed change (if any)**: are not useful (as they…	Steven Deleu	EORTC
N/A	Comment: "should not be" **Proposed change (if any)**: should be not	Steven Deleu	EORTC
N/A	No comments.	Christian Dittrich	ESMO