Anonymisation of Clinical Reports for Publication Under EMA Policy 0070

(Product Name: Zinbryta, Sponsor: Biogen)

Lukasz Kniola, Biogen, UK

EMA Technical Anonymisation Group (TAG) Meeting

29-30 November 2017

Submission Summary

- 11 phase I-III trials in Multiple Sclerosis
- Run between 2005 and 2014
- 2500+ unique patients across submission (some participating in more than one trial)

- 20 documents
- ~ 30k pages
- ~7.5k pages include one or more redactions of any kind (1 in 4 pages)

Direct and Quasi Identifiers

Direct identifiers	De-identification techniques
Subject IDs	Redacted
SAE IDs	Redacted

Quasi identifiers	De-identification techniques
Site IDs	Redacted
Age, Birth date	Redacted
Gender	Kept
Race	Redacted
Country	Redacted when presented in relation to individual subject or site.
Baseline body weight	Redacted
Subject's profession	Redacted
Visit, assessment, event, finding dates	Year kept, month and day redacted in calendar dates. Relative dates (number of days since first dose) were kept.

Other Data (Sensitive Information)

- Details and descriptions considered sensitive (HIV status, hepatitis status, gynaecologic history, psychiatric history, suicide attempts, illicit drug use, alcohol abuse) – carefully reviewed.
- Redacted only where details were highly unique (potentially known to a plausible attacker) or where successful linking to study participants could have severe consequences for the affected individuals.

Other Data (Verbatim Terms)

- Verbatim terms, including uncoded diagnoses, procedures, and uncoded substance names were thoroughly reviewed
- When presented in relation to individual subjects (e.g. in narratives, subject listings), coded terms were retained unless they or their combinations were deemed unique or sensitive.
- Where verbatim terms were similar or identical to coded terms, the same criteria were applied.
- Terms describing visible or unique characteristics or could reveal location or timing were selectively redacted.

Other Data (References and Comments)

- References to family, other persons, and place names were redacted.
- Investigator comments may contain geographic details, calendar dates, sensitive data, context specific to individual subjects. Those were carefully reviewed and selectively redacted to maximize data utility.

Redaction Technique

- External Vendor used
- Fields recognised as identifiers
- Rules applied consistently
- No flexibility in applying rules on a subject-by-subject basis

- Additional, manual review to identify sensitive information, family relationship, etc. (supported by programmatic search of the document, using pattern recognition, string search)

Quantifying Risk – Assumptions

- Data from Clinical Database used as basis for calculations.
- Maximum Prosecutor risk across all subjects in submission.
  Equivalence classes based on the population of submission (some subjects took part in more than one trial).
- Pre-defined threshold = 0.09.
- ”All or nothing ” rules – the available tool only allowed uniform application of each rule.
  Either all values of certain type were redacted or none of them.
  e.g. Age could be always redacted or always retained, but combination – on a subject-by-subject basis – was not possible.

Quantifying Risk – Calculations

- The probability of a re-identification attempt for public disclosure = 100%
- QI used for calculations: Site, age (+birth date), gender, race, country, baseline body weight, subject's profession, dates
- Iterative assessment to test all combinations of QIs.
- High variance of QI values, in combination with "all-or-nothing" rules, resulted in most QIs having to be redacted
- Only Gender and Year-part of dates were deemed safe to be retained.
- All other QIs were uniformly redacted.

Inference of values by cross-referencing multiple summaries

- Additional assessment established the need for selective redaction of frequency values in cases where unique events made it possible to collate subjects' details and those sets of details could be linked to <6 subjects in the population reported.

Term	Sex:Male	Sex:Female
XYZ	1	0
ABC	14	26
DEF		

Term	Age: <40	Age: >=40
XYZ	0	1
ABC	30	10
DEF		

Term	Region: NA	Region: EU
XYZ	0	1
ABC	21	19
DEF	15	5

Term:XYZ linked with 40+ year old Male in EU

If less than six 40+ year old males in EU across population then row XYZ redacted in one or more tables

Data Utility

- Aggregate summaries and analyses have the most scientific value and remained largely unmodified. Therefore, the results and conclusions of the study with regards to safety and efficacy findings on population and sub-population levels retain their value and utility.
- Narratives were selectively redacted. Details such as age, race, and country needed to be removed but the remaining information (AE, MH, CM) was largely retained (with the exception of sensitive information and references to family or location).

Possible Improvements

- Extracting level of detail for all subjects present in the reports (some may not have all QIs linked in text).
- Flexible application of rules.
  ”All-or-nothing” rules mean that if only a few subjects have unique values of any given QI, all instances of that QI for all subejcts need to be redacted.
  If only certain subjects could have selected QIs redacted it would allow for the less unique subjects (with lower individual risk) to retain more information.
  It would require manual application of those rules which was deemed impractical, due to the volume of submission.

- Ultimately, greated understanding of processes and availability of tools will allow for a shift from data redaction to data perturbation.

Comments received from EMA on submitted proposal package

- Requests for additional details and explanation in the AnR, where language was not clear
- Clarification on redaction of protected personal data of individuals other than trial subjects (it was not feasible to obtain consent from coordinating investigator's signatories as well as principal investigators)
- Request for rationale for removal of:
  - appendix (no copyright)
  - abnormal laboratory value listing (removed in line with Section 2.2 of the Policy)Instances of redacting the table/listing instead of removing
  - log listing (redaction applied, instead of removal)

Questions and Discussion