PhUSE 2017

Paper PP18

Approach to EMA Policy 0070 - From Data Science Perspective

Jorine Putter, Grünenthal GmbH, Aachen, Germany

ABSTRACT

With ever changing regulatory requirements, we data scientists must become increasingly agile, not only in our thinking towards how data should be collected, prepared, and reported, but also with "future-proofing" our patients' privacy from attacks stemming from future technological advancements. Our approaches need to be even more pragmatic and risk based when our business models may be widely spread between in-house products/compounds and in-licensed products/compounds during various phases of the lifecycle of drug development. This paper presents the cross functional approach Grünenthal is considering in this ever-changing landscape. I will discuss which factors are important to consider, outlining various approaches available to implement. In addition, I propose the approach we intend to employ, justifying the decision after thorough evaluation of the pros and cons of the relevant factors.

INTRODUCTION

The European Medicines Agency (EMA) policy on the publication of clinical data for medicinal products for human use was developed by the EMA, in accordance with Article 80 of Regulation (EC) No 726/2004. Policy 0070 was adopted by the EMA Management Board on 2nd October 2014 and subsequently published on the EMA website. Policy 0070 is composed of two phases. Phase 1 entered into force on 1st January 2015 and pertains to publication of clinical reports only. Phase 2, which will be implemented at a later stage, pertains to the publishing of individual patient data (IPD). Clinical reports and IPD are collectively referred to as "clinical data". [1]

To publish clinical reports, Policy 0700 requires application of anonymization techniques to clinical data to protect patients' privacy. For the purposes of this paper, I will focus on factors and approaches to consider for the implementation of Phase 1 of Policy 0070. For further information on the full scope, implementation of Policy 0070 and any further information related to Policy 0070 please consult the EMA website.

POLICY 0070 APPLIED TO DATA COLLECTION

Following the Study Data Tabulation Model (SDTM) v1.3 [2], data is categorized into the following main types of variables:

- Identifier – identifies the study, subject of the observation, the domain, and the sequence number of the record
- Topic – specifies the focus of the observation (such as the name of a test)
- Timing – describes the timing of the observation (such as start date and end date of an event)
- Qualifier – includes additional illustrative text, or numeric values that describe the results or additional traits of the observation (such as units or descriptive adjectives)

When considering data anonymization, two types of data identifiers are of primary interest: direct identifiers and quasi identifiers. According to the EMA Policy 0070 guidance [1] document:

1. *Direct identifiers* are elements that permit direct recognition or communication with the corresponding individuals. Direct identifiers generally do not have data utility, except for the patient ID. Direct identifiers include the patient's name, email, phone number, signature and full address.
2. *Quasi identifiers* are variables representing an individual's background information that can indirectly identify individuals. Unlike direct identifiers, information from quasi identifiers increases the scientific usefulness of the information published. Geographical location is an important variable since clinical practice can vary from country to country, impacting the outcome of the analysis. Relative dates (e.g. date of birth, start date of adverse event, etc.) relating to individual patients are also important due to the potential impact on the outcome of the trial. Patient level demographic information such as sex, age, race, ethnicity, height and weight can be confounders and therefore of critical scientific utility.

According to the guidelines provided by the PhUSE working group [3], quasi identifiers can be further split into two levels.

1. Quasi identifier level 1 – Information that is not likely to change over time, be visible and available in other sources. Typically, this includes demographic information such as sex, age at baseline, country, BMI, etc.
2. Quasi identifier level 2 – Longitudinal information is data that is likely to change over time, e.g. measurements, events, etc.

For sponsors to ensure that clinical reports are made public with minimized risk to the privacy of their patients, considerable thought must be given to what data is collected. Limiting the collection of data to what is required instead of what is "nice to have", not only helps data scientists minimize noise and allows focus on what the data is really telling us, but is critical to protecting our patients' privacy.

POLICY 0070 APPLIED TO DATA PREPARATION

The current data preparation within clinical trials typically flows as illustrated in Figure 1.

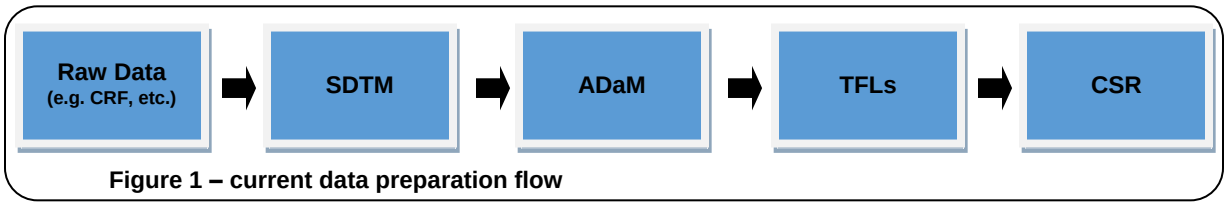

Figure 1 – current data preparation flow

The current data preparation within clinical trials that falls within the scope of EMA Policy 0070 Part 1 typically flows as illustrated in Figure 2. In this model, the anonymized CSR is prepared retrospectively.

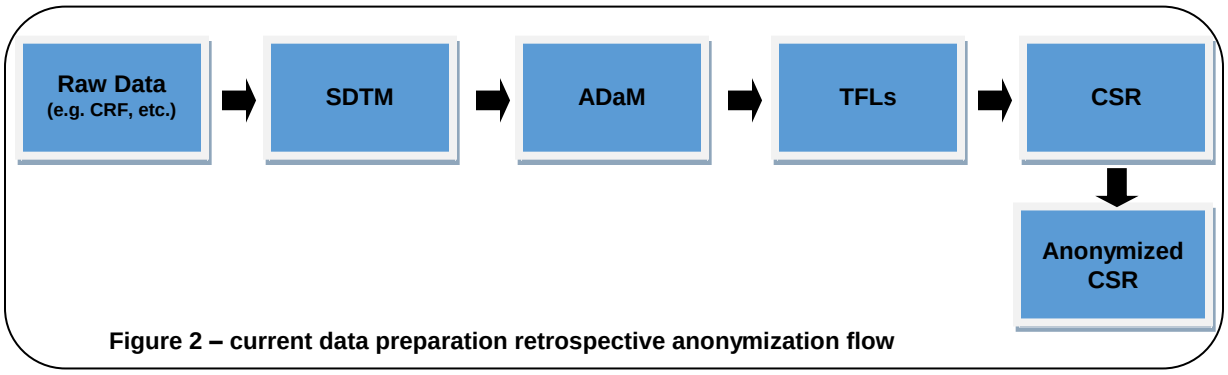

Figure 2 – current data preparation retrospective anonymization flow

An alternative scenario for data preparation within clinical trials that also falls within the scope of EMA Policy 0070 is illustrated in Figure 3. In this proposed model, the anonymized CSR is prepared prospectively at the first instance of processing collected data.

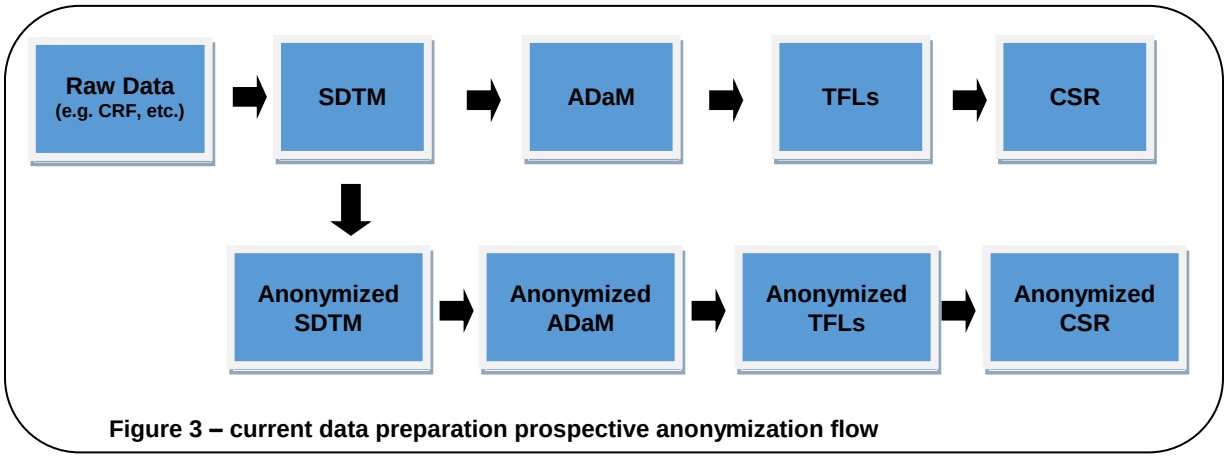

Figure 3 – current data preparation prospective anonymization flow

Sponsors should decide whether a retrospective, prospective, or hybrid approach fits them best. All approaches have benefits and drawbacks. When investigating which approach to follow, consider the following factors as a starting point.

- Size of your clinical program
- Maturity of your data preparation processes
- Agility of your data preparation tools
- Resource availability (FTE and/or costs)
- Maturity of regulations
- Knowledge within your organization
- Agility to alter anonymization methods

For sponsors to ensure that clinical reports are made public with minimized risk to the privacy of their patients, considerable thought must be given to how data should be prepared. Choosing the appropriate method should allow for sponsors to have the ability to meet regulations, but most importantly to protect our patients' privacy adequately, and in the most cost-effective use of sponsor resources.

POLICY 0070 APPLIED TO DATA REPORTING

What and how data is reported, is more important than ever before. When sponsors are planning their statistical analysis and what will ultimately become part of the TFLs for a CSR, the principles of EMA Policy 0070 should be kept in mind with regards to information being shared. Consider the following simple question: How will we handle a report that shows individual patient level information, either in the body or footnote of a table or figure? One aspect to consider is whether there is improved data utility in describing/identifying the specific patient, i.e. Subject ID, actual age, race, ethnicity, etc. If added value exists, then the sponsor should apply anonymization to that information in the anonymized CSR. If there is no additional analytical value, then the sponsor might consider refraining from showing individual patient level information in the body or footnote of a table or figure in the first instance. In another potentially deidentifying instance, sponsors might consider to use generalization techniques. This means replacement of a value by a range, i.e. instead of displaying the absolute minimum and maximum weight consider showing <55kg, >120kg or 50-55kg, 120-130kg.

PATIENT NARRATIVES / CASE NARRATIVES

The handling of patient narratives within the industry is a very hot topic currently and has been since the launch of EMA Policy 0070. The debate continues whether patient narratives should be included in the scope of the EMA Policy 0070, and if so, how they should be handled. For the purposes of this paper I will not go into the details of considerations for patient narratives, the only information I would like to bring to the attention of sponsors are that within this year at least two anonymization packages have been accepted, as per the EMA Clinical Data Portal [6], where patient narratives have been completely redacted. The patient narratives contain a large amount of sensitive information, resulting in a high probability of subject reidentification, hence complete redaction was accepted by the EMA. However, according to the EMA guidelines, sponsors are advised not to redact patient narratives entirely, but instead, apply adequate anonymization techniques. If, in exceptional cases, the entire patient narrative needs to be redacted to ensure anonymization, i.e. all identifiers (direct and indirect) need to be redacted, it has to be clearly justified in the anonymization report.

For sponsors to ensure that clinical reports are made public with minimized risk to the privacy of their patients, upfront thought must be given to what and how data is reported.

ANONYMIZATION PROCESS

As this is not entirely new to the industry, the industry might not be in its "infancy" years but it's certainly not much further along than "toddler" years. Within other industries anonymization is much further along, and the Pharmaceutical industry can certainly learn from the other industries.

There are several working groups within the Pharmaceutical industry (PhUSE, DIA (Drug Information Association), EFSPI (European Federation of Statisticians in the Pharmaceutical Industry), PSI (Statistician in the Pharmaceutical Industry), etc.), however the challenge is ensuring connectivity between these groups. During the finalization of the CSR, several functions (e.g. Biostatistics, Medical Writing, Clinical, etc.) must work closely together. Furthermore, an anonymized CSR requires input from almost all functional areas taking part in clinical development. Given the integral nature of data to the entire process, functions which prepare, analyze, report and interpret data are core contributors.

To walk you through the anonymization process, let's consider that the following example extract needs to be anonymized as per the EMA Policy 0070.

EXAMPLE EXTRACT

Double storey beige coloured house, with a blue tiled roof, chimney, five round windows and a grey door. Located at 1 Utility street.

Anonymization typically starts with determining your risk of re-identification assessment, either qualitatively or quantitatively. To determine your risk of re-identification quantitively, you need to determine your applicable population. There are several ways of doing this (e.g. prevalence in the world/regions where a sponsor's trial(s) has been conducted, epidemiology data, sample fractioning by utilizing public trial registries, pooling of several clinical trials performed on the same disease by the same company).

The following information is available for Utopia, where the trial was conducted:

TABLE 1: UTOPIA POPULATION

Category	Sub-category 1	Sub-category 2	Sub-category 3	Sub-category 4	Sub-category 5	N
House						150
House	Double Storey					50
House	Beige Colour					30
House	Tiled Roof					30
House	Blue Roof					40
House	Chimney					20
House	Round Windows					20
House	Five Windows					50
House	Grey Door					30
House	Beige Double Storey	Blue Tiled Roof	Chimney	Five Round Windows	Grey Door	1
House	Double Storey	Tiled Roof	Chimney	Five Round Windows		5
House	Tiled Roof	Chimney				11
House	Tiled Roof					30
House	Double Storey	Roof	Chimney	Several Windows	Door	12

Therefore, when performing a quantitative risk re-identification assessment and using the population information about where the trial was conducted (i.e. Table 1), the risk of re-identification assessment can be performed as below.

TABLE 2: IDENTIFY PPD (PERSONALLY PROTECTED DATA) PRESENT IN EXAMPLE EXTRACT

Identifier	Type
Address	Direct
Double storey	Quasi 1
House color	Quasi 2
Roof type	Quasi 1
Roof color	Quasi 2
House has chimney	Quasi 1
Shape of windows	Quasi 1
Number of windows	Quasi 1
Color of door	Quasi 2

Calculate the risk of re-identification, firstly on an individual quasi identifier basis and then on composite quasi identifiers. The risk is calculated as: risk = 1 / N. Below are the 1) Individual and 2) Composite risk results.

TABLE 3: INDIVIDUAL QUASI IDENTIFIER RISK

Quasi Identifier	N*	Risk
Double storey	50	2%
House color	30	3%
Roof type	30	3%
Roof color	40	3%
House has chimney	20	5%
Shape of windows	20	5%
Number of windows	50	2%
Color of door	30	3%

***** - numbers obtained from Table 1

TABLE 4: COMPOSITE QUASI IDENTIFIERS RISK

Composite Quasi Identifiers	N*	Risk
Double storey x House color x Roof type x Roof color x House has chimney x shape of windows x number of windows x Color of door	1	100%
Double storey x Roof type x House has chimney x shape of windows x number of windows	5	20%
Roof type x House has chimney	11	9.1%
Roof type	30	3.3%
Double storey x Roof type x House has chimney x number of windows x House has door	12	8.3%

***** - numbers obtained from Table 1

CHOOSE AN ACCEPTABLE RISK OF RE-IDENTIFICATION THRESHOLD

The EMA suggests 9% as an acceptable risk level, however this depends on several factors of which could include indication, rarity of disease and acceptable risk for a sponsor.

For the purposes of the example extract, the threshold is set at 9%.

ANONYMIZE IDENTIFIED PPD

To anonymize PPD as per EMA Policy 0070, the desired effect can be achieved with either redaction, the removal of text by blackening it out, or anonymization, the pseudo masking/alteration/generalization of the underlying data.

DIRECT IDENTIFIERS

For direct identifiers it's rather straight forward, either they must be redacted or pseudo masked, however to pseudo mask them requires altering the data, which might not be an option depending on the agility of a sponsors' data preparation and reporting tools.

For the purposes of the example extract, there is one direct identifier (Table 1) *ADDRESS* which will be redacted.

QUASI IDENTIFIERS

For quasi identifiers, first determine their own and composite risk of re-identification values (Table 3 and Table 4 respectively). When these values are higher than the acceptable risk of re-identification, they should be redacted or anonymized. However, anonymization requires altering the data, which might not be an option depending on the agility of a sponsors' data preparation and reporting tools. Some anonymization techniques can be found in various articles, e.g. *PhUSE Data Transparency Group* [3], *Anonymising and sharing individual patient data* [4], etc. This is an iterative process, until all single and/or composite risk of re-identification values are below the acceptable risk of re-identification.

For the purposes of the example extract, there are no individual quasi identifier that exceeds the acceptable threshold of 9%. However, several composite quasi identifiers do exceed the acceptable threshold of 9%. By redacting certain quasi identifiers (*HOUSE COLOR, ROOF COLOR, SHAPE OF WINDOWS* and *COLOR OF DOOR*) and anonymizing *NUMBER OF WINDOWS* (replacing *"five"* with *"several"*) an acceptable risk of re-identification factor of 8.3% is achieved.

DATA UTILITY

Once the anonymization of PPD is deemed as complete, a data utility assessment should be completed. The sponsor's aim should be to reduce the risk of re-identification to the acceptable level and retain as much data utility as possible. Ideally, the data utility assessment should be performed quantitively. There are various techniques to apply, one that I found very useful is described in the article *An Enhanced Utility-Driven Data Anonymization Method* [5].

"FUTURE PROOFING"

A final consideration when applying Policy 0700 and anonymizing the data involves future-proofing the resulting information. The discussion to this point considers the static technology available to today's data scientists. However, the resulting data will live forever and technological advances in data processing, data mining, or statistical procedures may increase the risk for re-identification in the future. With the ease of use and availability of these types of software, sponsors must stay on top of what developments are out there and adapt their approaches keeping this risk in mind. It is almost impossible to ensure an entirely future proof anonymization, but due diligence is required, with the lack thereof sponsors could adversely impact the future risk to the privacy of their patients.

CONCLUSION

Given that knowledge is power and data leads to knowledge, the sharing of data is of utmost importance. Not only does the availability of data allow us data scientists the possibility to create more efficient models, but ultimately this is to the direct benefit of our patients as they can receive their desired medicines faster and more cost effectively with far greater transparency. The current trend in the industry is to be reactive to making data publicly available, however to ensure compliance and minimize the risk to our patients' privacy, there needs to be a mind shift to proactive thinking and planning and not treating this very important aspect as an afterthought.

Many sponsors might be of the opinion that a prospective approach could lead to submission delays, in part they might be correct. However depending on the specific factors surrounding a sponsor's data preparation processes and tools, a carefully thought-out and implemented prospective approach to handling policies, such as the EMA Policy 0070, could not only aid submission timelines and policy compliance, but there could be a tremendous cost, time and FTE saving compared to a retrospective approach.

REFERENCES

[1] External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use v1.3, European Medicines Agency policy on publication of clinical data for medicinal products for human use (EMA/90915/2016).

[2] Study Data Tabulation Model (SDTM) v1.3, CDISC. Study Data Tabulation Model (SDTM) v1.3.

[3] PhUSE Data Transparency Group, PhUSE De-Identification Working Group, "De-Identification Standards for CDISC SDTM 3.2," 2015.

[4] Anonymising and sharing individual patient data, Khaled El Emam, Sam Rodgers, Bradley Malin. Anonymising and sharing individual patient data. BMJ 2015; 350: h1139.

[5] An Enhanced Utility-Driven Data, Stuart Morton, Malika Mahoui, P. Joseph Gibson & Saidaiah Yechuri. TRANSACTIONS ON DATA PRIVACY 5 (2012) 469 - 503.

[6] EMA Clinical Data Portal, European Medicines Agency Portal for Clinical Data.

RECOMMENDED READING

1. Protection of personal data in clinical documents - a model approach, TransCelerate Biopharm Inc., Clinical Data Transparency 2017.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jorine Putter
Grünenthal GmbH
Zieglerstr. 6
Aachen / 52078
Germany
Email: Jorine.putter@grunenthal.com