PhUSE 2017

Paper RG02

Evaluation of Re-identification Risk for Anonymized Clinical Documents

Parveen Kumar, GCE Solutions Inc., Chandigarh, India
Rajan Sareen, GCE Solutions Inc., Chandigarh, India

ABSTRACT

Clinical data sharing for data transparency has potential to strengthen academic research, the practice of medicine and the integrity of clinical trial systems. This topic is becoming popular nowadays amongst leading pharmaceutical companies, and it raises concerns around the potential leak of personal health information of patients participating in clinical trials. The European Medicines Agency (EMA) has published phase 1 of their policy 0070 around requirements for anonymization of clinical documents for all studies submitted to EMA for market authorization, signaling a commitment towards data sharing, and increasing the urgency for addressing risk of potential re-identification. This paper will be discussing different scenarios in anonymization of clinical documents, implications of re-identification risk and the need for automation.

INTRODUCTION

It is not only EMA that is moving towards clinical data sharing but other associations/big pharmaceutical companies are trying to make data sharing mandatory. Pharmaceutical companies have started submitting anonymized clinical documents to the EMA with redaction of any text meeting the criteria of personal information or Commercially Confidential Information (CCI). Clinical documents in this paper is meant for all the reports that needs to be submitted under phase 1 policy of Policy 0070. In July 2013, Pharmaceutical Researchers and Manufacturers of America (PhRMA) & European Federation of Pharmaceutical Industries and Associations (EFPIA) member companies demonstrated commitment to share complete Clinical Study Reports (CSRs) along with Individual Patient Data (IPD) with qualified scientific and medical researchers, as necessary to conduct legitimate research. Recently in June 2017, International Committee of Medical Journal Editors (ICMJE) have introduced a requirement for data sharing statements for clinical trials before publishing any articles in journals abiding to ICMJE. However, in this paper, we shall be mostly interested in discussing about risk assessment methods for submission of clinical documents under EMA Policy 0070.

Data redaction may lead to compromised data utility; hence the EMA is interested to approach to alternative techniques for data anonymization, including replacement, generalization, date offsetting, etc., with a more clearly established method for calculating quantitative risk of re-identification. This needs lots of efforts and resources to complete it manually. Industry needs an automated process of anonymizing clinical documents like study reports and Clinical Narratives. Anonymization methodologies must include a way of measuring re-identification risk and have a repeatable process to follow. Submissions done at EMA under phase 1 of policy 0070 have qualitative risk assessments. This paper will be discussing different ways of risk assessment methods suggested in literature for anonymized clinical documents.

DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the author. The interpretations of standards and procedures contained in this paper are those of the author and are not necessarily correct. Any views and recommendations stated within this paper are those of the author, and they do not represent the position of their employer.

DEFINITIONS

The current terminology in data sharing and data protection may be confusing: some terms are often used interchangeably. For clarity, the following definitions have been used in this paper:

ANONYMIZATION/DE-IDENTIFICATION

The process of rendering data into a form which does not identify an individual and where re-identification is not likely to take place. Anonymization and de-identification can be used interchangeably except in case of anonymization there are no keys back to original data. For our own understanding, anonymization is used in case of clinical documents, where there are no key datasets back to original documents, and de-identification is used in case of IPD.

DIRECT AND QUASI-IDENTIFIERS

The first step in de-identification process is to identify variables which can directly or indirectly lead to identification of subjects participated in study. Direct identifiers are the variables that can on their own identify subjects within the dataset. In the clinical trial scenario, all types of IDs (like Subject ID, SAE ID, sample numbers, etc.). Quasi-Identifiers are the variables which helps to identify a subject within the dataset, when used in combination with other quasi-identifiers. These include age, sex, country as well as dates, unique events, etc. Quasi-identifiers are subsequently used in calculating the risk of re-identification of the dataset. PhUSE de-identification standard document is a good guide to understand selection of identifiers in standard dataset. Same variables can also be considered identifiers in related clinical documents.

ASSESSMENT OF ANONYMIZATION

EMA has given two scenarios for establishing whether clinical documents are adequately anonymized or not. Sponsor needs to select one of the two scenarios and explain its application to submission documents.

FULFILMENT OF CRITERIA FOR ANONYMIZATION

It needs to be demonstrated that after anonymization, it is no longer possible to:

- Singling out: Is it still possible to single out an individual? Direct identifiers like subject IDs, names, etc. can lead to singling out a particular individual.
- Linkability: Is it still possible to link records relating to an individual and identify that individual? Ability to link to data records of same individual or group of individuals and deduce their identity. Ability to linking one Quasi-identifier like age, sex, race, date of birth, medical history etc. with another can lead to identification of a particular individual.
- Inference: Can information be inferred concerning an individual? It is a possibility to deduce few attributes of an individual with certain probability.

It is clear that it is very difficult to meet the above criteria and maintain data utility at the same time. Only possibility is that if CSRs do not contain any information on direct and quasi-identifiers then above criteria can be met. Even TransCelerate states that “given the limited experience applying this option, it is difficult to provide guidance around implementation until more is understood.”

RISK ASSESSMENT

Whenever a proposal does not meet the criteria mentioned in last section, demonstration of effective anonymization shall be done by evaluating re-identification risks. Measuring the risk of re-identification involves:

- An appropriate risk metric
- A suitable threshold
- Actual measurement of risk

The choice of risk metric depends on the context of data release. As Policy 0070 lead to public data sharing, we will discuss the risk assessment for intention of public release of anonymized CSRs. Threshold advised by EMA is 0.09, if risk is assessed analytically and threshold shall be low risk, if assessed using qualitative approach. There are two ways to assess actual risk i.e. qualitative and quantitative. It was acknowledged by EMA that during early implementation of phase 1 of policy 0070, risk assessment could be mostly qualitative but sponsors are encouraged to move to quantitative methods as soon as they are in position to do so.

It is also important to take into account that with an advancement in technology (e,g, data mining), there will be a greater availability of data which may further increase the possibility of database linkage and the risk of data re-identification. Thus, anonymization should not be regarded as one-off exercise and the attending risks should be reassessed regularly by data controllers based on advances in technology and availability of advanced techniques of evaluation of risk of re-identification.

QUALITATIVE RISK ASSESSMENT

A method that assesses the risk of re-identification based on the characteristics of the source data and not on IPD. The approach uses a qualitative scale (eg, high, moderate, low) for the assessment of risk. Factors to be considered for qualitative risk assessment are largely:

- Number of direct and quasi-identifiers within a report
- Size and nature of population of disease studied (i.e. rare or common disease, pediatric etc.)
- Number of participants in study
- Number of study centers and distribution of centers across countries

Table 1 illustrates the impact of these factors on overall qualitative risk. For example, in scenario 1, if there are only few identifiers reported within report out of high number of participants in multicenter study then risk of re-identification could be low. Similarly, scenario 1 and scenario 2 depicts the situation where risk could be moderate or high. In general, the risk of re-identification increases with small studies or studies with only few sites. Similarly, rare patient populations may also increase the risk of re-identification.

Table 1: Qualitative risk assessment factors

Factor	Scenario 1	Scenario 2	Scenario 3
Number of identifiers within report	Low	High	High
Size of population of disease studied	High	Moderate	High
Number of participants	High	Moderate	High
Number and distribution of study centers	High	Moderate	Low
Risk	Low	Moderate	High

Risk assessment will mostly be based on quasi-identifiers as all direct identifiers are supposed to be either redacted or recoded and shall have no contribution to risk. Any direct identifier missed would lead to re-identification of that individual and risk of re-identification will be 1. Thus, direct identifiers will be redacted or recoded in CSRs in all situations.

Once general risk assessment is done, sponsor should specifically work on anonymization technique to be applied to reduce the high or moderate risk to low risk. Low risk is considered to be the threshold for all types of qualitative risk assessment. If risk is high, then anonymization should be strict and all the direct and quasi-identifiers shall be either redacted/anonymized. If initial risk calculated is moderate, then anonymization of direct and key quasi-identifiers may serve the purpose to reduce the risk level to low. Sponsor needs to explain very clearly in anonymization report regarding initial risk, measures taken to reduce it, and final risk after anonymization. “Protection of personal data in clinical documents – A model approach” by Transcelerate has explained the approach to anonymize clinical documents in case of qualitative risk assessment.

QUANTITATIVE RISK ASSESSMENT

Transcelerate defined Quantitative risk as “A method that analyzes the data itself to measure the risk and how to best de-identify the data by establishing a probability. The approach measures the risk as a numerical value.”

It has now become common to use automated tools to redact/anonymize clinical documents but this could lead to two types of risk.

- One is risk of automation tool missing to annotate any direct or indirect identifier that is present in document and this could lead to re-identification of an individual.
- Second is due to actual values or anonymized values in document that adds to risk of re-identification. For example, say, age is generalized from actual value of 24 years to a range of 20-25 years, even then the anonymized value have some risk of re-identification associated with it.

RISK DUE TO TEXT DE-IDENTIFICATION TOOLS

Scaiano et al. (2016) proposed a unified framework for evaluation in terms of the probability of re-identification when medical text is de-identified using automated tools. This paper compares this unified framework with other approaches available in literature and states benefit of current framework. Basic assumption around tool to be used is that all of the information pertaining to an individual trial participant can be extracted as a unit and treated as a separate virtual document for the purpose of evaluation.

This framework primarily considers recall (r) value which is defined as number of correct annotations (tp) by tool, divided by sum of number of correct annotations (tp) and number of missed annotations (fn) i.e r= tp/(tp+fn) where,
tp are those which shall be identified by tool and are identified

fn are identifiers leaked or missed to be identified by tool.

Framework worked around different assumptions for direct and quasi-identifiers. It was assumed that any leak in direct identifier and any leak in at least two quasi-identifiers would lead to re-identification. Probability of interest here is a conditional probability of re-identification, if there is an attempt to identify a leak that appeared in a document i.e.

Pr(reid, attempt, leak, appears)= Pr(reid/attempt,leak,appears) × Pr(attempt/leak,appears) × Pr(leak/appears) × P(appears)

As per Scaiano’s paper, the probability for

<u>Direct identifiers</u>:

$$\begin{aligned} &Pr(\text{reid}, \text{attempt}, \text{leak}, \text{appears}) \\ &\quad = Pr(attempt|leak, appears) \\ &\quad\quad \times \left(1 - \prod_{\{i|r_i \geqslant 0.9\}} (1 - h \times w_i(1 - r_i)) \prod_{\{i|r_i < 0.9\}} (1 - w_i(1 - r_i))\right) \end{aligned} \quad \text{……………………} \quad (1)$$

where h is the probability a leaked identifier value is successfully hiding in plain sight i.e. the probability that an adversary cannot distinguish if the value is original one that leaked or one that was resynthesized, w_i is proportion of subjects where direct identifier i, appeared out of total number of subjects, r_i is recall value for identifier i.

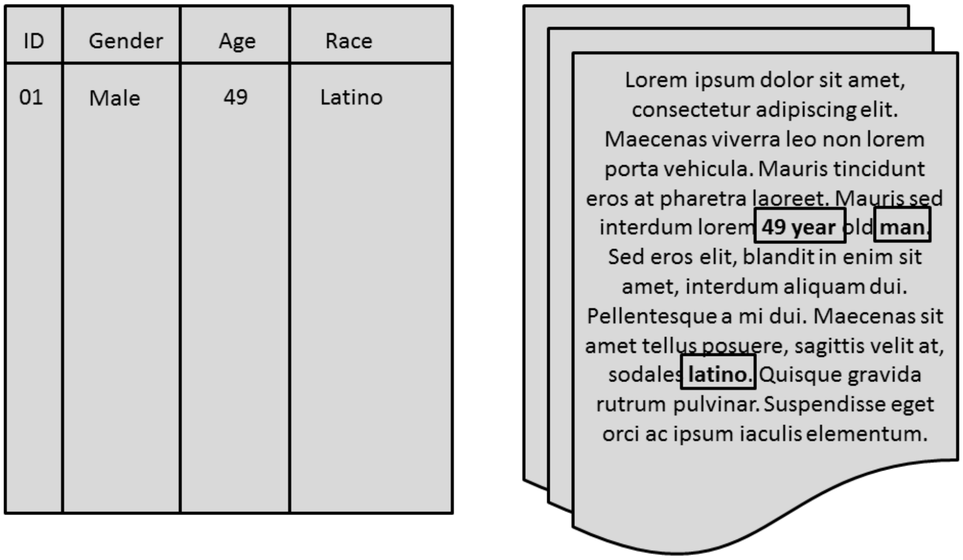
Quasi-identifiers

$$\begin{aligned}&Pr(reid, attempt, leak, appears) = Pr(attempt|leak, appears)\\ &\times Pr(X \geqslant 2 \text{ if } r_q \geqslant 0.7,\ \text{or } Y \geqslant 2 \text{ if } r_q < 0.7)\\ &\text{for}\\ &X \sim B(n_q, h(1-(r_q)^m)), Y \sim B(n_q,(1-(r_q)^m))\end{aligned} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2)}$$

Where h is the probability a leaked identifier value is successfully hiding in plain sight i.e. the probability that an adversary cannot distinguish if the value is original one that leaked or one that was resynthesized, r_q is recall for quasi-identifier q, n_q is average number of distinct quasi-identifier values per document, m is number of times, on average, that a quasi-identifier values in document is repeated and B(a,b) is a binomial distribution with 'a' trials and 'b' probability of success.

RISK BASED ON GENERALISED QUASI-IDENTIFIERS IN CLINICAL DOCUMENTS

Generally, structured IPD is available to the trial sponsor but not to researcher who receives de-identified CSR. Clinical documents contain information of subjects from IPD data. Following is a random example of correspondence between IPD and CSR for a study participant (example taken from appendix of Scaiano 2016).



ID	Gender	Age	Race
01	Male	49	Latino

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas viverra leo non lorem porta vehicula. Mauris tincidunt eros at pharetra laoreet. Mauris sed interdum lorem **49 year** old **man.** Sed eros elit, blandit in enim sit amet, interdum aliquam dui. Pellentesque a mi dui. Maecenas sit amet tellus posuere, sagittis velit at, sodales **latino.** Quisque gravida rutrum pulvinar. Suspendisse eget orci ac ipsum iaculis elementum.

Structured IPD

CSR Pseudo-Documents

If quasi-identifiers in IPD are de-identified using generalization or other techniques, then de-identified values from IPD can be used to replace original values in CSR. In place of redaction, generalized values will be able to maintain the document utility up to some extent. Risk of re-identification of IPD based on defined quasi-identifiers can be calculated for each subject. EMA says "the probability of re-identification of a record in a dataset is 1 divided by the frequency of trial participants with same category/value of a set of the quasi identifiers (group size)". Calculation is explained in more detail by Kniola (2016). Once we have risk calculated for each subject, risk of re-identification for clinical document containing information on all or few subjects from IPD can be derived.

DIFFERENT SCENARIOS FOR QUANTITATIVE RISK ASSESSMENT

Now, we know that there could be two different types of risk in clinical documents; 1) risk due to automated tool or manual review missed on annotating any direct or quasi-identifier which should have been annotated. This may lead to re-identification of subjects 2) risk due to generalization of quasi-identifiers and not redacting them completely.

There is a possibility of different scenarios for any submission under phase 1 of Policy 0070 and accordingly the probability mentioned in above section will be joined or modified as mentioned below:

PUBLIC SHARING OF REDACTED DOCUMENTS

Redacted clinical documents will be the documents where direct and quasi-identifiers are redacted completely and any risk of re-identification will be based on leak in annotation of valid identifier.

As per EMA Policy 0070, clinical documents will be shared on public platform. Hence, probability of attempt by adversary is considered to be 1 and in equations (1) and (2), Pr(attempt/leak, appears) will be considered equal to 1. Probability of attempt to re-identify any individual will be calculated if there is non-public disclosure of data and there are different ways to calculate it as described by Scaiano 2016 or Kniola 2016. However, this is out of scope for this paper.

Sponsors are currently using redaction as the method to anonymize all the direct and key quasi-identifiers in documents. There could be multiple reasons for it as below:

- Redaction may be the most practicable option when data analysis and document writing is complete.
- Most of the tools in market to anonymize documents are proficient in doing redaction only.
- Redaction itself is very time consuming if it has to be performed manually. If other techniques has to be applied then it will be more resource consuming.

If redaction is the only method used for anonymizing documents for EMA policy 0070 then equations (1) and (2) will be modified to remove HIPS factor and Probability of attempt will be described as below:

For direct identifiers:

Pr(reid/leak,appears) = $1 - \prod_{i} (1 - w_i(1 - r_i))$ ……………………………………………….(3)

For quasi-identifiers:

Pr(reid/leak,appears) = $Pr(X \geqslant 2) \quad \text{for } X \sim B(n_q, 1 - (r_q)^m)$ ………………………….(4)

REPLACEMENT USING DE-IDENTIFIED IPD

Once IPD data is de-identified using methods such as generalize or other transformations then same transformations could be applied to CSR as well. For example, if age is replaced with five years interval in IPD then all instances of age in CSR would be resynthesized to five years interval matching with IPD.

If additional direct identifiers like Name, email addresses, phone, etc. are redacted, direct identifiers for subjects participating in study will be recoded as per de-identified data. This will modify the risk of re-identification due to quasi-identifiers. If there is any leak, Quasi-identifiers will still benefit from *hiding in plain sight*. For quasi-identifiers, probability of re-identification due to leak will not be zero and would lead to a change in equation (2) as below:

$$\left(\left(\Pr(\text{reid}|\text{catch}) \times \left(R_q\right)^M\right) + \Pr\left(X \geq 2 \text{ if } r_q \geq 0.7, \text{or } Y \geq 2 \text{ if } r_q < 0.7\right)\right)$$

$$\text{for } X \sim B\left(N_q, h\left(1 - \left(R_q\right)^M\right)\right), Y \sim B\left(N_q, \left(1 - \left(R_q\right)^M\right)\right)$$

Where

$R_q \sim \mathrm{N}\left(r_q, \sqrt{r_q(1 - r_q)/s_q}\right), N_q \sim \mathrm{Pois}(n_q), M \sim \mathrm{Pois}(m)$

And Pr(reid/catch) will be risk calculated on IPD data. There is a chance that all quasi-identifiers do not appear in document while risk is for all quasi-identifiers. This an err on conservative side.

Direct identifiers will also benefit from *hiding in plain sight* and probability of re-identification due to leak will remain same as equation (1).

RISK BASED ON IPD DATA ONLY FOR ANONYMIZED CLINICAL DOCUMENTS

CSRs and other clinical documents will always have an IPD data linked to it. As discussed earlier, it is a good approach to anonymize clinical documents using the de-identified IPD data i.e. to replace all direct and quasi-identifiers in clinical document with de-identified values in de-identified IPD data. There are two ways to anonymize clinical documents i.e. using automated tool or manually. This will be suggested here that probability of risk of re-identification due to missed annotation or leak can be ignored if we eradicate all the possibilities of missing to annotate any direct or quasi-identifiers. Even if automated tool is used, there will be manual review to check if anonymization is done as expected or not. Following are the rationales for making the above point:

- It is expected that at least no direct identifiers are leaked in clinical documents. One direct identifier missed to be anonymized means that there is definite re-identification in case of public disclosure of documents. Hence, it is not affordable to miss anonymization of even a single direct identifier for a single patient.
- Clinical documents that are supposed to be submitted for phase 1 of Policy 0070 are not supposed to be very lengthy as appendices containing individual patient data listings are not in scope for phase 1. Medical writers or someone with good exposure to ICH E3 will have an idea that which sections will possibly have PPD information. This way any document which is anonymized using a tool can be quickly reviewed for any missed annotation by an experienced person. This will ensure that there is no missed annotation for direct or quasi-identifiers.

If risk due to missed annotation is removed, then the risk of re-identification will depend only on risk calculated based on IPD data. It can be done in two ways as below:

Conservative approach: The most appropriate way to measure the risk of re-identification for an entire IPD dataset, in the context of public disclosure, is through the maximum risk, which corresponds to the maximum probability of re-identification across all records. It will be nothing but maximum of all the individual risk calculated for each subject (refer Kniola 2016 for more details). This approach is conservative in nature because there is a chance that not all the subjects with all the quasi-identifiers from IPD are mentioned in clinical document. For example, if there are 100 subjects in a dataset and only 10 of it are mentioned in clinical documents then risk calculated on 100 subjects is assigned to risk of re-identification in clinical document as well. There is err in conservative way. But suppose if the subject with maximum risk is NNN and there is detail around this subject in clinical document with all the quasi-identifiers with de-id data then there will be no err in risk for clinical documents. For example, Table 2 is referring to Kniola (2016) for calculating risk for each record and assuming only two quasi-identifiers in a study i.e. Sex and Age:

Table 2: Dataset with equivalence class sizes and subject level risk of re-id

USUBJID	SEX	AGE	Equiv. Class (Size)	Re-Id Risk
CT1/101	M	26	A(3)	0.33
CT1/102	F	26	B(2)	0.5
CT1/103	M	26	A(3)	0.33
CT1/104	F	26	B(2)	0.5
CT1/105	F	29	C(2)	0.5
CT1/106	F	28	D(2)	0.5
CT1/107	M	26	A(3)	0.33
CT1/108	F	27	E(1)	1
CT1/109	F	28	D(2)	0.5
CT1/110	F	29	C(2)	0.5

Overall risk for above IPD data will be maximum of all individual risk values i.e. 1. We can assign this overall risk to clinical document containing information in form of text for all or few of these subjects from IPD. If subjects 108 is mentioned somewhere in clinical with their sex and age referred then this risk assigned to clinical document has no err otherwise there will be some err on the conservative side.

Less conservative approach: If automated tool or manually we are able to find the list of subjects that are mentioned in clinical documents with list of all quasi-identifiers linked to each subject then we can subset the IPD risk for only those subjects. For example, if we know that only subjects 101 and 102 have appeared in clinical document with information around their age and sex then we know that re-id risk associated to these two records in IPD is 0.33 and 0.5, respectively. Hence, re-id risk for clinical document will be max of these two i.e. 0.5.

There could be one more situation that not all the quasi-identifiers have appeared in clinical document. For example, if only subjects 101 and 102 have appeared with age information only then re-id risk will be re-calculated for each record based on age as quasi-identifier only. It is re-calculated in Table 2 and now re-id risk will be 0.2 in clinical documents with only subject 101 and 102 appearing in it with age information only.

Table 3: Dataset with equivalence class sizes and subject level risk of re-id

USUBJID	AGE	Equiv. Class (Size)	Re-Id Risk
CT1/101	26	A(5)	0.2
CT1/102	26	A(5)	0.2
CT1/103	26	A(5)	0.2
CT1/104	26	A(5)	0.2
CT1/105	29	B(2)	0.5
CT1/106	28	C(2)	0.5
CT1/107	26	A(5)	0.2
CT1/108	27	D(1)	1
CT1/109	28	C(2)	0.5
CT1/110	29	B(2)	0.5

AUTOMATION

Clinical documents are type of structured documents predefined templates. Automation for anonymizing clinical documents could be very effective and may lead to saving of lots of efforts and resources with increasing efficiency of anonymized documents. Automation for different scenarios as below is needed to meet the requirements of phase 1 of Policy 0070:

- Automated process to detect direct and quasi-identifiers in clinical documents and redact it as per regulatory requirement. For example, replacing direct and quasi-identifiers with PPD with white color font in blue box. This would be an ideal situation for legacy studies where analysis is already done and documents are available without access to IPD data.
- Automation process to de-identify IPD data and then replace all the appearing direct and quasi-identifiers in clinical documents with de-id values from IPD. This method will ensure good quality of documents since none of the text is getting redacted but replaced with generalized, recoded, and randomized values.

Although automated process makes the de-id process fast but manual review plays a vital role to ensure that there is nothing missed to annotate or something unnecessarily annotated by the automated tool. Automated tool needs a manual override or addition to clinical documents to ensure that CCI is taken care of manually. Artificial Intelligence to automated process could also be useful here so that tool is learning on its own that which sections or parts of documents have what type of direct and/or quasi-identifiers.

CONCLUSION

Anonymizing clinical documents for submission under phase 1 of Policy 0070 could be challenging and a time-consuming process. Automation tools are there in market but they are not handling risk assessment effectively. This paper has discussed multiple ways to address risk assessment problem for anonymized clinical documents under different scenarios. Most of the sponsors are following redaction approach to anonymize documents and qualitative approach of risk assessment. EMA has encouraged in their external guidance document to not only redact the identifiers but to replace these identifiers with de-id values and use quantitative approach for risk assessment. This approach not only ensures that the risk is managed well but also ensures that anonymization is not too excessive to affect data utility.

This paper suggests a method where IPD data shall be available along with clinical documents and be used for anonymizing clinical documents and quantitative risk assessment. This will also prepare sponsors for phase 2 of Policy 0070 where de-identified IPD data needs to be available for submission. Limitation of this approach could be that 1) it will be time consuming, if automated tool is not available to replace de-id values from IPD in clinical documents, 2) Rigorous manual review of anonymized clinical documents will be needed so as to ensure that tool has performed as expected.

REFERENCES

European Medicines Agency. (2014). EMA/240810/2013 - *Publication of clinical data for medicinal products for human use*. http://www.ema.europa.eu.

European Medicines Agency. (2017). EMA/90915/2016 - *External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use*. http://www.ema.europa.eu.

Scaiano (2016) et al. *A unified framework for evaluating the risk of re-identification of text de-identification tools*. Journal of Biomedical Informatics.

TransCelerate (2016) – *Protection of personal data in clinical documents – A model approach*.

CONTACT INFORMATION

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

Parveen Kumar
GCE Solutions Inc.
327, Bestech Tower A, Mohali
Punjab, India - 160062
Work Phone: 011-41069686
Email: parveen.kumar@gcesolutions.com