PhUSE 2016

Paper DH09

Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach

Lukasz Kniola, Biogen, Maidenhead, UK

ABSTRACT

There is an increasing obligation, demand and expectation from regulatory bodies, independent researchers as well as in-house secondary research needs to share clinical trial results and data. Consequently, the concerns for patients' anonymity must be addressed. In response Biogen formed a Data Sharing group whose purpose is, among other tasks, to manage and perform a thorough, defendable calculation of the risk of individual patients being re-identified. This paper will focus on the metrics and methodology employed by our group to quantify the risk of re-identification and managing that risk by finding the right level of de-identification of patient data depending on the context of the data release. It will also discuss assumptions we used for selecting the adequate risk level thresholds and ways to achieve it. Finally, it will consider challenges of data de-identification in the current environment (like EMA Policy 0070) and going forward.

INTRODUCTION

There are numerous reasons why clinical trial data may need to be shared. From in-house re-use, to allowing external researchers perform additional analyses, to regulatory requirements like EMA Policy 0043 and EMA Policy 0070, any type of secondary research, where data is used for purposes other than those specified in the original protocol and not covered by the informed consent may require a level of de-identification. Depending on who the designated recipient is and how the data is shared different measures and controls are available and different criteria need to be met.

It is important to stress that the aim of de-identification is to reduce the risk of re-identification to an acceptable level while retaining as much data utility as possible. This is in contrast to removing the risk altogether which would inevitably deem the data unusable.

Calculating the risk of re-identification should ensure that the right balance is struck between how much information can be shared and how much should be suppressed to preserve patients' identity safety.

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 positions of their employer.

CONTEXT OF DATA RELEASE

It is important to differentiate between several contexts of data release.

Sharing data with known researchers, under strict contracts, through secure means like sandbox will ensure that the process is safe and the risks involved are very low. On the other hand risk is very high in the context of a public release.

CONTRACT

Before data is shared a number of decisions need to be made.

Where the process is bounded by a contract, the sponsor may enforce requirements and restrictions. The most common of them include:

- Prohibition on re-identification attempts.
- Prohibition on attempting to contact any of the subjects in the data set.
- Audit requirement allowing to conduct spot checks to ensure compliance with the agreement, or requirement for regular third-party audits.
- Requirement to pass on controls to any other party the data is subsequently shared with.

DATA PREPARATION

An important decision is whether to release all available data related to a clinical trial or only select datasets and variables in those datasets which support the object of the data sharing exercise. Being able to reduce the scope may result in less de-identification being required. A careful consideration is needed and it is vital to consider future uses of the de-identified data. Once a streamlined data set is shared it may not be possible to release a more comprehensive set at a later stage without putting the identity of the subjects at risk.

Similarly, if clinical trial data had been previously released to public domain in any form and scope, it may aid an attacker in targeting another set of the same data released in more secure ways.

In general, when the same data is shared in different contexts it is advisable to consider the relation between those various releases.

PUBLIC RELEASE

When data is released to the public domain, there is no control over how it will be used and a potential attacker wanting to access the data can do so with little effort. In this scenario the most conservative assumptions are required as it is impossible to assess the motivations of the attacker and the level of knowledge and tools they may possess and use in ways which may be harmful to the sponsor.

IDENTIFIERS IN THE DATA

The first step in the de-identification process is to find variables which can directly or indirectly identify subjects in the data set. An invaluable resource to aid in this task is the PhUSE De-identification Standard for CDISC SDTM document. It can be used to help find identifying variables and rules to apply to those variables.

DIRECT IDENTIFIERS

These are variables that can on their own identify subjects within the data set. In the clinical trial scenario these are all types of IDs (like Subject ID, SAE ID, sample numbers, etc.).

With the exception of Subject ID, those variables have typically no analytical value and can be removed from the data set. Subject ID needs to be pseudonymised in a way that any link between original and new random IDs is broken.

QUASI-IDENTIFIERS

These are variables which can help identify a subject within the data set when used in combination with other quasi-identifiers. These include age, sex, country as well as dates, unique events, etc. For full list of quasi-identifiers and rules that can be applied to them, refer to the PhUSE De-identification Standard document.

Quasi-identifiers are subsequently used in calculating the risk of re-identification of the data set.

PROBABILITY OF RE-IDENTIFICATION ATTEMPT

The literature identifies three categories of re-identification attempts: deliberate attempt, inadvertent (spontaneous) attempt and data breach. Each of these categories depends on the context of data release and controls the sponsor has in place.

Probability of re-identification attempt should only be considered when data is shared under contractual agreement, when recipient is known and measures can be taken to ensure the data is used under agreed rules.

In the context of public data release, the probability of attempt is set to 1 (or 100%) as the sponsor has no control over who will access the data and how it will be used.

One such scenario is data release under EMA Policy 0070. Since information is made publicly available the probability of attempt is set to 1.

DELIBERATE ATTEMPT

In deliberate attempt an attacker targets either a specific subject or any subject in the data set. Its probability will be a function of controls over the use of data as well as the means and the motivation of the potential attacker.

If the sponsor has an established relationship with the data recipient it is unlikely that an attack will be attempted as it would severely impact the relationship. When data is shared with a researcher unknown to the sponsor then reputation may be considered.

A small research unit is not likely to have the necessary resources to perform an attack while big organizations may present a higher risk of attempting to use the data for their advantage.

Data shared through a secure portal, with access restrictions and reduced ability to list data at patient level, greatly reduces the risk. Releasing the actual data via sFTP, email or thumb drive reduces the level of control on who will use the data.

Finally, the sensibility of the data shared and the potential harm in an event of re-identification needs to be carefully considered.

The sponsor will need to take all those factors into consideration in order to assess the deliberate attempt risk. It will typically take a value between 0.1 (or lower) for scenarios where data is shared with a trusted researcher using secure means and 0.6 (or even 0.75) when requestor isn't known to the sponsor or data is considered to be sensible.

For more details see literature (Institute of Medicine, 2015), (El Emam, 2013), (Canadian Institute for Health Information, 2010).

INADVERTENT RE-IDENTIFICATION

An inadvertent re-identification occurs when a researcher working with the data recognizes someone in the data set. The risk of such scenario is equivalent to the probability of the data recipient knowing a subject within the data set well enough to identify them based on the information available.

In order to estimate this risk in the context of clinical trials two values need to be considered:

- Average number of friends people tend to have – according to the leading anthropologist Robin Dunbar this is about 150.
- Prevalence of the disease that the clinical trial is targeting. This is defined as number of cases of the disease divided by the population of the region in question.

Using those values, the probability of acquaintance can be calculated using the following formula:

$$Pr(acquaintance)=1-(1-(Disease\ Cases/Population))^{150}$$

Using example prevalence and population values in Table 1, risk of acquaintance can be calculated as per Table 2.

Table 1. Prevalence of MS and general population

	MS Cases	Population (age 15-65)	Overall Population
US	400K	212M	317M
Worldwide	2.3M	4.7B	7.2B

Table 2. Calculated risk of acquaintance

	Population (age 15-65)	Overall Population
US	0.25	0.17
Worldwide	0.07	0.05

DATA BREACH

The last category is a data breach. This occurs when the data recipient loses the data (the physical medium on which data was stored is lost or stolen) or when recipient's security systems are breached.

Literature quotes the HIMSS (Healthcare Information and Management Systems Society) survey which estimates the average breach rate at 0.27. This value can be applied to scenarios where data is provided to the recipient in the form of raw patient level information.

It should be noted that the risk of a data breach is significantly reduced when access to data is only allowed through a portal with relevant security measures. In such cases the sponsor may choose to revise the risk level.

OVERALL RISK OF AN ATTEMPT

Once all three risk levels are calculated the worst case should be selected for further calculations.

For example in a scenario where data is shared with a researcher with whom a good relationship is established (low risk of deliberate attempt) and for a rare disease (low risk of inadvertent re-identification) but the data is transferred on a thumb-drive (high risk of breach) then it is the last of the three that will likely have the highest value.

If similar data is shared using a secure portal but the recipient is an unknown researcher, the deliberate attempt may pose the highest risk.

Overall:

$$Pr(attempt)=max(\ Pr(deliberate\ attack),\ Pr(acquaintance),\ Pr(breach)\ )$$

PROBABILITY OF SUCCESSFUL RE-IDENTIFICATION OF A SUBJECT

CALCULATING RISK FOR EACH SUBJECT – STUDY POPULATION VS WIDER POPULATION

The next step is to calculate the probability of re-identification for each subject in the data set shared should an attack be attempted.

The following is an example dataset with two quasi-identifiers – sex and age.

Table 3. Example dataset

USUBJID	SEX	AGE
CT1/101	M	26
CT1/102	F	28
CT1/103	F	31
CT1/104	M	29
CT1/105	F	28
CT1/106	M	30
CT1/107	M	29
CT1/108	F	32
CT1/109	M	29
CT1/110	F	31

Let's first sort all subjects sharing the same quasi-identifiers values into groups called equivalence classes. The probability of re-identification of an individual record is then the reciprocal of its class size:

Table 4. Dataset with equivalence class sizes and record level risk of re-id

USUBJID	SEX	AGE	Equiv. Class (Size)	Re-Id risk
CT1/101	M	26	A (1)	1
CT1/102	F	28	B (2)	0.5
CT1/103	F	31	C (2)	0.5
CT1/104	M	29	D (3)	0.33
CT1/105	F	28	B (2)	0.5
CT1/106	M	30	E (1)	1
CT1/107	M	29	D (3)	0.33
CT1/108	F	32	F (1)	1
CT1/109	M	29	D (3)	0.33
CT1/110	F	31	C (2)	0.5

Individual probabilities for each record can then be used to calculate the overall risk across the dataset.

The above is correct if the aim is to calculate the risk of re-identification only taking into account the dataset population. In other words, it is irrelevant how many people share similar characteristics in the general population since we are only focusing on the sample of the general population that is our clinical trial.
This means however that – especially in clinical trials with few subjects – the individual risks will be high and the final values of risk will not meet the thresholds. This can in turn lead to excessive de-identification and loss of data utility.

To calculate the realistic re-identification risk it would be required to have detailed census-like data on different diseases in multiple regions. Knowing numbers of patients with the same disease within the same region sharing the quasi-identifier values would drastically reduce the risks as they would be based on general population rather than the trial sample.

In reality, obtaining and maintaining such a vast data set would be impractical and very costly. Since this information would likely need to be combined using many sources, using it would require complicated statistical models and a level of estimation and assumption.
What pharmaceutical companies do have at hand is the data from other trials they have run which can be used to estimate the risks.
As an example, if the aggregate database of subjects in similar studies contained 12 male subjects at the age of 26, then the Re-Id risk for USUBJID=CT1/101 would be 1/12 = 0.083, etc.

Table 5. Applying record level risk of re-id based on similar trials dataset

SEX	AGE	Equiv. Class Size
M	26	**12**
F	28	**32**

USUBJID	SEX	AGE	Equiv. Class (Size)	Re-Id risk
CT1/101	M	26	A (12)	0.083
CT1/102	F	28	B (32)	0.031
CT1/103	F	31	C (27)	0.037
CT1/104	M	29	D (11)	0.091
CT1/105	F	28	B (32)	0.031
CT1/106	M	30	E (15)	0.067
CT1/107	M	29	D (11)	0.091
CT1/108	F	32	F (4)	0.250
CT1/109	M	29	D (11)	0.091
CT1/110	F	31	C (27)	0.037

Although the risks calculated using this approach will be higher than those based on general population, they are much more straight forward to apply and easier to defend if methods are questioned. Crucially, they are more conservative than general population risks but will typically be significantly lower than risks calculated solely on the

study population. In the example, CT1/101 has a risk of re-identification of 1 within the study population, but this value is reduced to 0.083 when similar studies are taken into account.
Across the data set this will mean less de-identification efforts and more data usability.

CALCULATING RISK FOR DATA SET – AVERAGE VS MAXIMUM

Once individual risks for each record are known, the overall risk of re-identification of the dataset in the event of an attempt can be calculated.
The two metrics typically calculated are the average risk and the maximum risk.
The maximum risk will take the highest value of all individual risks for all records across the dataset.
The average risk will be the mean of all individual risks across the dataset.
In the example sets, for study population only, these will be:

$$Pr_{max}(re\text{-}id \mid attempt) = max\ (1, 0.5, 0.5, 0.33, 0.5, 1, 0.33, 1, 0.33, 0.5) = 1$$

$$Pr_{avg}(re\text{-}id \mid attempt) = (1+0.5+0.5+0.33+0.5+1+0.33+1+0.33+0.5) / 10 = 0.6$$

while risks across similar trials:

$$Pr_{max}(re\text{-}id \mid attempt) = max\ (0.083, 0.031, 0.037, 0.091, 0.031, 0.067, 0.091, 0.250, 0.091, 0.037) = 0.25$$

$$Pr_{avg}(re\text{-}id \mid attempt) = (0.083+0.031+0.037+0.091+0.031+0.067+0.091+0.250+0.091+0.037) / 10 = 0.081$$

The maximum risk is by definition no lower than the average risk and is therefore much stricter.
When data is released to the public domain and there is no control over who receives the data and how they choose to use it, it has to be presumed that an attack will be attempted. And because the likely target of an attack will be the records with highest probability of re-identification, the maximum risk has to be considered for those scenarios.
When data is shared under contractual restrictions and the recipient of the data is known, then the average risk is the metric to choose. However it is still useful to calculate maximum risks in those scenarios. It helps to understand how varied the equivalence class sizes are and avoid leaving unique records in the data set.

K-ANONYMITY AND UNIQUENESS

Although the value of the average risk may be low enough to meet the threshold, this on its own might not be sufficient to deem the data safe for sharing. One additional metric to utilize is *k*-anonymity, where *k* is the smallest equivalence class we are happy to leave in the shared dataset. For 2-anonymity this means that there are at least two records that share any combination of values of quasi-identifiers, for 5-anonymity, at least five records, etc.
One way to implement this metric is to make sure that after all de-identification techniques have been applied, the *k*-anonymity condition is also met. This way not only is the average risk acceptable but also no less than *k* records share the same characteristics in the data set.
A less strict approach is to calculate the percentage of records not meeting the *k*-anonymity criterion. Under certain conditions, instead of de-identifying the data set to the point where *k*-anonymity is achieved across all records, it may be permissible to allow a small number of records (for example, <1%) to remain unique (< *k* similar records in data set). Calculating the proportion of unique records can also aid in the iterative process of de-identification as it can show how close to meeting all criteria the data is and identify equivalence classes which need to be taken care of.
For sensitive data or for public release a careful consideration is required when using similar trials to assess *k*-anonymity. In those cases, even if overall risk for the data set is derived based on similar trials, it may be sensible to assess *k*-anonymity using only study population (only records in the data set) to avoid leaving unique records in the data set.

COMBINING IT ALL TOGETHER

Once all risk elements have been calculated, it is time to combine them to produce a comprehensive set of metrics describing the level of de-identification and risks of de-identification of the entire data set.
The first step is to derive the overall risk of re-identification which is the product of risk of attempt and risk of successful re-identification of records in the set:

$$Pr(re\text{-}id) = Pr(re\text{-}id \mid attempt) \times Pr(attempt)$$

This way average and maximum risk of data set re-identification can be calculated. The same formula will apply to study population and across similar trials. The following set of characteristics will be produced:

Table 6. Calculated metrics for the example dataset

	Pr_{max}(re-id)	**Pr_{avg}(re-id)**	**% of not *k*-anonymous records (*k*=2)**
Study population	1	0.6	33.3%
Similar trials	0.25	0.081	0%

THRESHOLDS AND RIGHT LEVEL OF DE-IDENTIFICATION

The three characteristics can subsequently be compared to the assumed thresholds to assess if re-identification is sufficient or if further data perturbation is required.

Setting adequate thresholds to be employed in re-identification risk assessment has been a topic of many discussions and publications. There are many precedencies supporting a range of values. On one end of the spectrum very permissive values as high as 0.33 can be found. These allow for very little data manipulation but are arguably not safe especially in the context of clinical trials. On the other end thresholds of <0.05 are used. These are very strict and typically require a great level of data perturbation which significantly reduces the usability of the de-identified data.

A threshold which is often met in the context of clinical trials and which offers a good balance between risk of data re-identification and data usability is a conservative value of 0.09. The same value is suggested as the preferred threshold for the EMA Policy 0070 submission and is becoming widely adopted by the clinical trials world.

For public release, the selected threshold will be applied to the maximum calculated risk. For controlled data sharing it will typically be applied to the average risk. In those scenarios it is ultimately the decision of the organization responsible for releasing the data. The maximum risk will always be higher and therefore safer, but using it may in some instances lead to unnecessary loss of data usability.

Using the values calculated previously in combination with different contexts of data release, below are some examples which illustrate that the same values may be sufficient in some scenarios and unacceptable in others.

Table 7. Risk calculations for different data release contexts

Context	Estimated risk of attempt in the context	Pr_{max}(re-id) / Threshold	Pr_{avg}(re-id) / Threshold	% of not *k*-anonymous records (*k*=2) / acceptable value	Result
Trusted recipient/ Secure portal/ Study population	0.1	0.1 / NA	0.06 / 0.09	33.% / 1%	**Not sufficient**
Trusted recipient/ Secure portal/ Similar Trials	0.1	0.025 / NA	0.008 / 0.09	0% / 1%	Sufficient
Trusted recipient/ Data on CD-ROM/ Study population	0.27	0.27 / NA	0.162 / 0.09	33% / 0%	**Not sufficient**
Trusted recipient/ Data on CD-ROM / Similar Trials	0.27	0.067 / NA	0.022 / 0.09	0% / 0%	Sufficient
Unknown recipient/ Secure portal/ Study population	0.5	0.5 / NA	0.3 / 0.09	33% / 0%	**Not sufficient**
Unknown recipient/ Secure portal/ Similar Trials	0.5	0.125 / NA	0.041 / 0.09	0% / 0%	Sufficient
Public release/ Study population	1	1 / 0.09	1 / NA	33% / 0%	**Not sufficient**
Public release/ Similar Trials	1	0.25 / 0.09	0.081 / NA	33% / 0% *	**Not sufficient**

*) For public release k-anonymity is calculated for study population only

Based on these examples it can be concluded that the context can significantly impact the estimated risk. If the control measures under which the data is released cannot be changed then the only practical solution to lower the overall risk is to undergo another iteration of data manipulation.

The example data set used in this paper only consists of ten records. It is therefore not surprising that using study population only results in re-identification risks being higher than the thresholds for all scenarios.

When risks are calculated using similar trials, all but the public release scenarios show that the risk of re-identification is below the threshold of 0.09. However, even in those scenarios the level of de-identification would not be sufficient since the proportion of records not meeting the *k*-anonymity criteria is too high (33.3%).

To lower the risk of re-identification when attempted and to address the *k*-anonymity issue, data would require further manipulation. In the example data set the practical options would be to categorize age values into groups. The below example shows this scenario and updated risk values:

Table 8. Updated dataset with categorized age and recalculated record level risks

USUBJID	SEX	AGE	AGE_{DE-ID}	Equiv. Class (Size)	Re-Id risk
CT1/101	M	26	21-30	X (5)	0.2
CT1/102	F	28	21-30	Y (2)	0.5
CT1/103	F	31	31-40	Z (3)	0.33
CT1/104	M	29	21-30	X (5)	0.2
CT1/105	F	28	21-30	Y (2)	0.5
CT1/106	M	30	21-30	X (5)	0.2
CT1/107	M	29	21-30	X (5)	0.2
CT1/108	F	32	31-40	Z (3)	0.33
CT1/109	M	29	21-30	X (5)	0.2
CT1/110	F	31	31-40	Z (3)	0.33

Table 9. Re-calculated metrics for the example dataset

	Pr_{max}(re-id)	Pr_{avg}(re-id)	% of not *k*-anonymous records (*k*=2)
Study population	0.5	0.3	0%

Table 10. Updated risk calculations for different data release contexts

Context	Estimated risk of attempt in the context	Pr_{max}(re-id) / Threshold	Pr_{avg}(re-id) / Threshold	% of not *k*-anonymous records (*k*=2) / acceptable value	Result
Trusted recipient/ Secure portal/ Study population	0.1	0.050 / NA	0.030 / 0.09	0% / 1%	Sufficient
Trusted recipient/ Data on CD-ROM/ Study population	0.27	0.135 / NA	0.081 / 0.09	0% / 0%	Sufficient
Unknown recipient/ Secure portal/ Study population	0.5	0.250 / NA	0.150 / 0.09	0% / 0%	**Not sufficient**
Public release/ Study population	1	0.500 / 0.09	0.300 / NA	33% / 0%	**Not sufficient**

Table 10 only shows values for study population calculations. Values for similar trials would be reduced even more than those for study population as equivalence groups would also be derived using categorized age groups. And so where for subject CT1-101 there were 12 equivalent records where SEX=M and AGE=26, across similar trials, after categorization the equivalence group will be made of records where SEX=M and AGE_{DE-ID}=21, 22…, 30 across similar trials.

DOCUMENTING ASSUMPTIONS, TECHNIQUES AND RESULTS

The final but crucial step of the risk assessment is documenting all assumptions, techniques used and calculations. De-identification is probabilistic and it is never possible to reduce the risk of re-identification to zero while retaining any data usability. A carefully done, well documented risk assessment can however show that we have done the due diligence and that every effort was made to minimize the risk of data being de-identified to an acceptable level.
It is not enough to show that we have done "something". Justification of methods and metrics used is equally important.
The following are the decisions required in the process of data de-identification and should be documented:

- Context of data release (contract, relationship with data recipient, means of providing data, sensitivity of data, prevalence of disease – common vs rare, public release).
    - In case of controlled release means and motivation to re-identify inform the risk of attack.
    - For public release special measures should be considered.
- Consideration of other releases of the same data set (actual and potential).

- Selection of direct and quasi-identifiers as well as techniques and rules applied to those in the effort to de-identify the data.
- Decision on whether to use only study population or other similar trials to calculate record level risks.
- Use of average or maximum risk across the data set.
- *k*-anonymity – select level of *k* and proportion allowed in the de-identified data (none or very low).
- Selection of thresholds and comparison of results.
- Data utility considerations.

EMA POLICY 0070

European Medicines Agency policy on publication of clinical data for medicinal products for human use became effective on 1 January 2015. Further details on the operational aspects of the policy were published in a guidance released in March 2016. The policy is composed of two phases. The first phase pertains to publication of clinical reports only and is active already. The second phase will involve the release patient-level data. It is not known at the time of writing this paper when it will be implemented.

The industry is currently on a steep learning curve to embrace the requirements and expectations of the policy. There are challenges ahead and discussions and experiences of the first submissions will lead to more clarity in process and risk assessment.

What transpires from the policy is the following:

- Only clinical reports and not the patient-level data (like listings) is the subject of the first phase.
- Narratives included in the reports are not exempt from the policy and therefore are subject to de-identification and publication.
- As this is a public data release – maximum risk across the data set is to be considered and compared against the threshold.
- Public release means that the risk of re-identification attempt is 1 (as described in the Context of Data Release paragraph).
- The threshold for maximum risk suggested in the policy is 0.09.

Ways to calculate risk of re-identification of data contained in a free-flowing text are being developed but it is clear that it is not as straight forward a process as it is when structured data sets are assessed. Information needs to be extracted from the text before risk can be measured. Once this is achieved, in principle, the methods described in this paper are applicable although there are some special considerations resulting from the nature of free-flowing text (e.g. number of quasi-identifiers listed in the text may differ from subject to subject which means that record-level risks may need to be calculated in a more dynamic manner). The initial wave of submissions is expected to use different methods of both anonymizing the reports and estimating the risk of re-identification. However, since EMA favours the quantitative approach, it is expected that well defined and documented processes and techniques will emerge to address this expectation.

CONCLUSION

Calculating the risk of re-identification of patient-level data using quantitative approach is the most comprehensive way to show due diligence in the process of data de-identification and ensure the security of the data being shared in relation to the context of the data release. Not only does it ensure that the risk is managed and minimized to acceptable level but it can also be a useful tool to show that the de-identification is not too excessive. This is important in the context of data utility.

Just like insufficient de-identification may lead to residual risk being high and data released not being safe, too aggressive de-identification is likely to compromise the usability of the data. It is therefore important to strike the right balance and using the quantitative approach helps to prove that the level of de-identification of the data is both sufficient and not too conservative and that the data is safe while retaining its usability.

REFERENCES

Canadian Institute for Health Information. (2010). *'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information.* Ottawa, ON: www.cihi.ca.

El Emam, K. (2013). *Guide to the De-Identification of Personal Health Information.* Boca Raton, FL: CRC Press.

El Emam, K., Arbuckle, L. (2013). *Anonymizing Health Data Case Studies and Methods to Get You Started.* Sebastopol, CA: O'Reilly Media.

Institute of Medicine. (2015). *Sharing Clinical Trial Data Maximizing Benefits, Minimizing Risk.* Washington, DC: The National Academies Press.

Pharmaceutical Users Software Exchange (PhUSE). De-identification standards for CDISC SDTM 3.2.

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. (2016). *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

CONTACT INFORMATION

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

Lukasz Kniola
Biogen Idec Ltd
Innovation House
70 Norden Road
Maidenhead, Berkshire SL6 4AY
UK
Email: lukasz.kniola@biogen.com