



Derive Value From Excellence ...

Evaluation of Re-identification Risk for Anonymized Clinical Documents



Agenda

- 1) Recent Background
- 2) EMA Policy 0070
- 3) Categorization of Clinical Data
- 4) Assessment of Anonymization
- 5) Criteria for Anonymization
- 6) Qualitative Risk
- 7) Quantitative Risk
 - a) Component 1
 - b) Component 2
- 8) Questions



RECENT BACKGROUND



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

2 October 2014
EMA/240810/2013

European Medicines Agency policy on publication of clinical data for medicinal products for human use



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

11 April 2017
EMA/90915/2016
Version 1.2 published 12 April 2017

External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use

Annals of Internal Medicine

EDITORIAL

Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors

The International Committee of Medical Journal Editors (ICMJE) believes there is an ethical obligation to responsibly share data generated by interventional clinical trials because trial participants have put themselves at risk. In January 2016 we published a proposal aimed at helping to create an environment in which the sharing of deidentified individual participant data becomes the norm. In response to our request for feedback we received many comments from individuals and groups (1). Some applauded the proposals while others expressed disappointment they did not more quickly create a commitment to data sharing. Many raised valid concerns regarding the feasibility of the proposed requirements, the necessary resources, the real or perceived risks to trial participants, and the need to protect the interests of patients and researchers.

It is encouraging that data sharing is already occurring in some settings. Over the past year, however, we have learned that the challenges are substantial and the requisite mechanisms are not in place to mandate universal data sharing at this time. Although many issues must be addressed for data sharing to become the norm, we remain committed to this goal.

Therefore, ICMJE will require the following as conditions of consideration for publication of a clinical trial report in our member journals:

1. As of 1 July 2018 manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement as described below.
2. Clinical trials that begin enrolling participants on or after 1 January 2019 must include a data sharing plan in the trial's registration. The ICMJE's policy regarding trial registration is explained at www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html. If the data sharing

plan of data sharing statements that would meet these requirements are in the Table.

These initial requirements do not yet mandate data sharing, but investigators should be aware that editors may take into consideration data sharing statements when making editorial decisions. These minimum requirements are intended to move the research enterprise closer to fulfilling our ethical obligation to participants. Some ICMJE member journals already maintain, or may choose to adopt, more stringent requirements for data sharing.

Sharing clinical trial data is one step in the process articulated by the World Health Organization (WHO) and other professional organizations as best practice for clinical trials: universal prospective registration; public disclosure of results from all clinical trials (including through journal publication); and data sharing. Although universal compliance with the requirement to prospectively register clinical trials has not yet been achieved and requires continued emphasis, we must work toward fulfilling the other steps of best practice as well—including data sharing.

As we move forward into this new norm where data are shared, greater understanding and collaboration among funders, ethics committees, journals, trialists, data analysts, participants, and others will be required. We are currently working with members of the research community to facilitate practical solutions to enable data sharing. The United States Office for Human Research Protections has indicated that provided the appropriate conditions are met by those receiving them, the sharing of deidentified individual participant data from clinical trials does not require separate consent from trial participants (2). Specific elements to enable data sharing statements that meet these requirements have been adopted at ClinicalTrials.gov (<https://pinfo.clinicaltrials.gov/definitions.html#shareData>). The WHO also supports the addition of such elements at the pri-

ClinicalStudy
DataRequest.com

Registered Users, Please Login

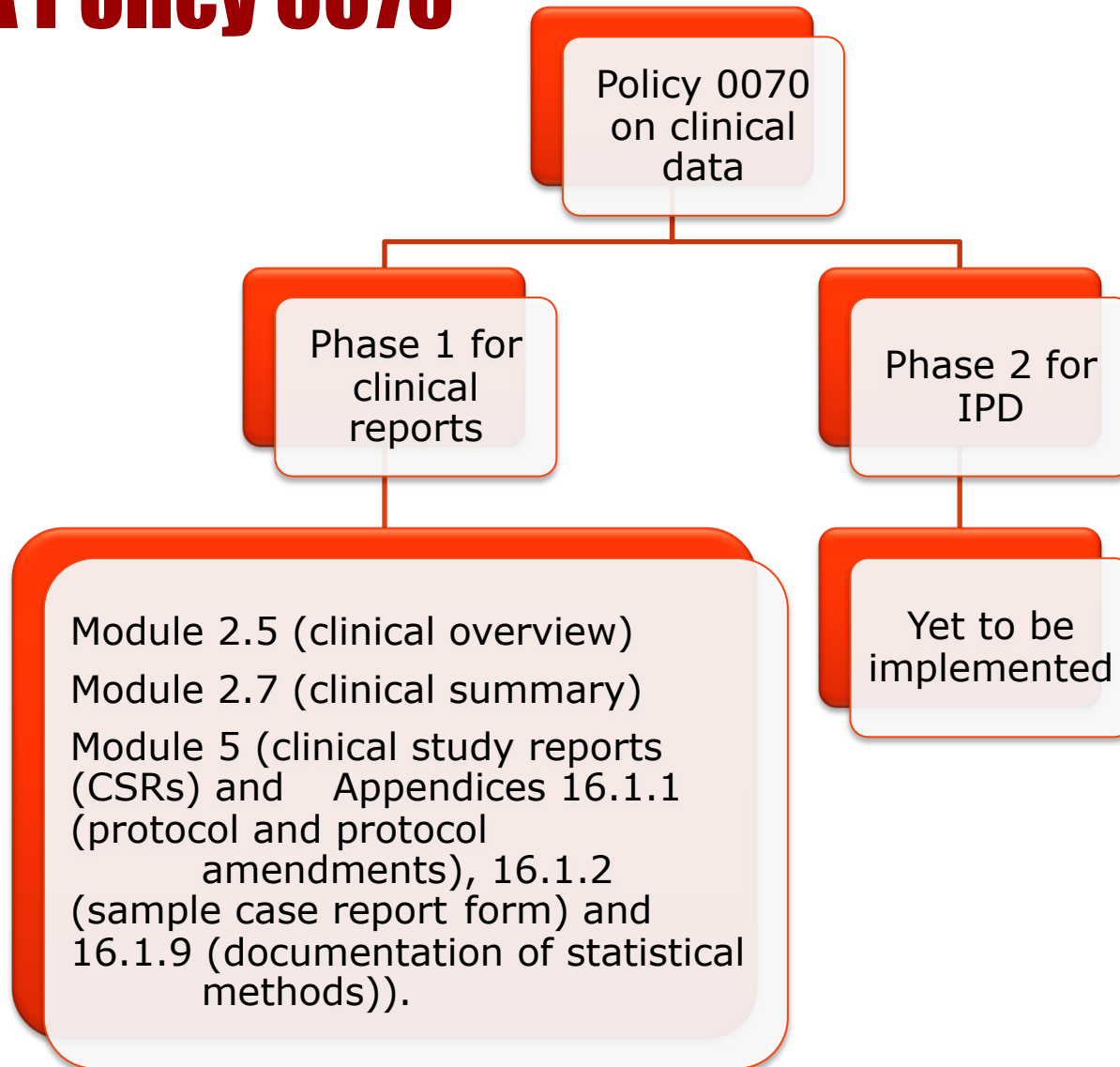
- HOME
- STUDY SPONSORS
- STEP BY STEP
- MY REQUESTS
- LOGIN OR CREATE AN ACCOUNT
- METRICS
- HELP

Get started

View »	View and submit »
You can view studies listed on this site before creating an account... »	After you create an account, you can select studies and submit a research proposal or enquiry... »



EMA Policy 0070



Key Note: *External Guidance recently updated to provide specific list of documents to be published under Phase 1.*



Categorization of Clinical Data

Quasi-Identifiers	Direct Identifiers:	Sensitive Information
Demographic: Race, Ethnicity, Date of Birth, Age, Sex, Country Longitudinal: Adverse Events (Serious): Death, Hospitalization; Start/End of AE – MHTERM – SITEID	USUBJID, INVNAME	Substance abuse, mental health disorders, HIV, reproductive health, sexually transmitted diseases – MHTERM (text) – medical history
Risk: The data may lead to identification of subjects participating in the study directly or indirectly. Per EU legislations, this personal data must be protected. Adversaries: Deliberate attempt, acquaintance, Data breach Remedy: Anonymization of clinical trial data to significantly reduce the risk for reidentification of subjects.		



Assessment of Anonymization

Methods to establish if data has been properly anonymized:

- 1) Demonstrate that after anonymization, it is no longer possible to:
 - Single out an individual from the rest of the data (“Singling Out”)
 - Link datasets and records related to an individual together and use the combined information to identify that individual (“Linkability”)
 - Infer an attribute of subject from other attribute with some significant probability (“Inference”)

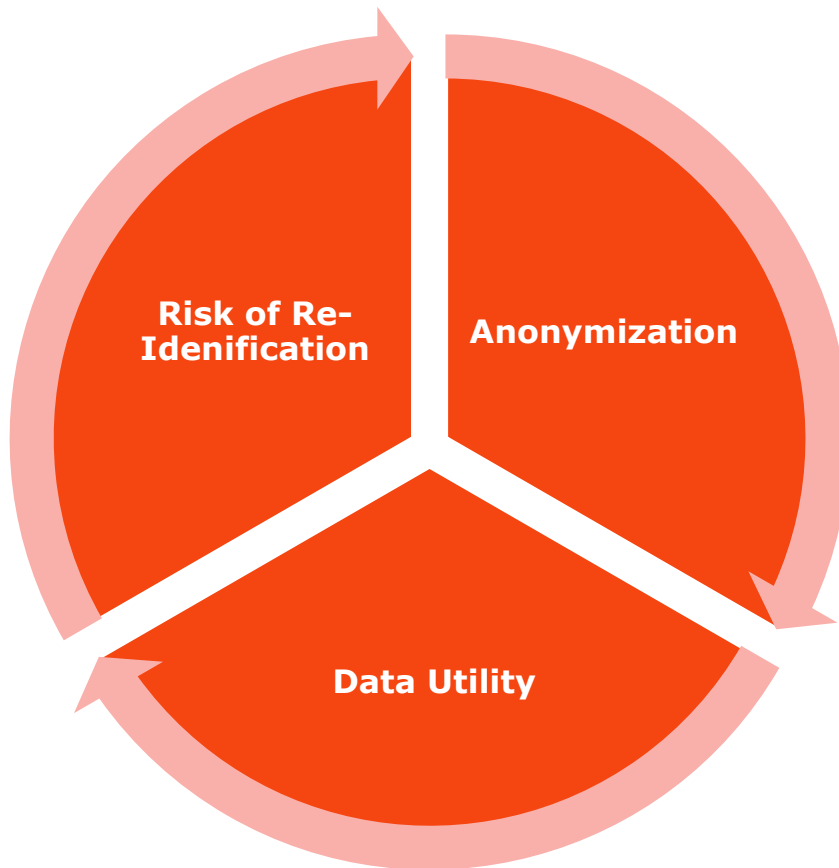
Note: It is clear that it is very difficult to meet the above criteria and maintain data utility at the same time. Only possibility is that if CSRs do not contain any information on direct and quasi-identifiers then above criteria can be met.

OR

- 2) Perform an analysis of re-identification risk, including:
 - An appropriate risk metric
 - A suitable/established threshold
 - An actual, calculated measurement of risk



Criteria for Anonymization

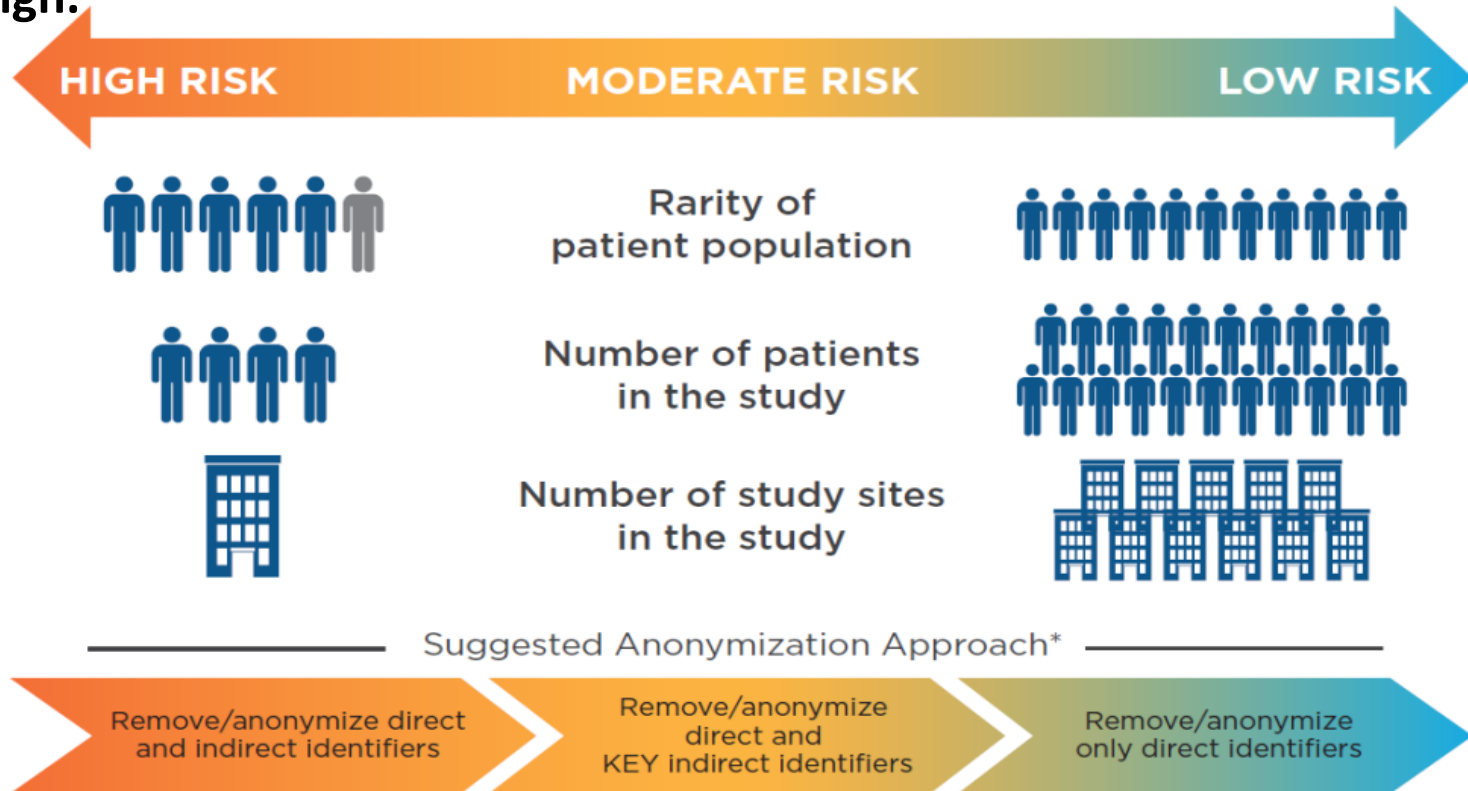


- 1) Meet established thresholds for risk of re-identification – for EMA: 0.09
- 2) Confirm anonymization methods do not impact data utility, i.e., generally defined as the ability to replicate analyses/results



Qualitative Risk Assessment

Approach measures risk as qualitative value i.e. low, medium or high.



* copied from Transcelerate guidance document

Other factors to be considered: Age range, race, number of identifiers within the data/report.



Quantitative Risk Assessment

Approach measures risk as numerical value...

Two components of risk of re-identification for de-identified data that recent literature [(Scaiano et al. (2016) & Kniola (2016))] has identified are:

A Unified Framework for Evaluating the Risk of Re-identification of Text De-identification Tools

*Martin Scaiano^{1,2}, Grant Middleton³, Luk Arbuckle⁴, Varada Kolhatkar^{1,2}
Liam Peyton¹, Moira Dowling⁵, Debbie S. Gipson^{5,6}, Khaled El Emam^{1,2,3,4*}*

¹School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

²Privacy Analytics Inc., Ottawa, Canada

³Pediatrics, University of Ottawa, Ottawa, Canada

⁴Children's Hospital of Eastern Ontario Research Institute, Ottawa, Canada

⁵Honest Broker Office, University of Michigan Medical School, Office of Research, Ann Arbor, United States of America

⁶Department of Pediatrics, University of Michigan, Ann Arbor, United States of America

- 1) Risk due to missing to annotate/redact any direct or Quasi-identifier which is planned to be annotated.

- 2) Risk due quasi-identifiers which are caught by process and transformed or left unchanged in clinical document.

PhUSE 2016

Paper DH09

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

Lukasz Kniola, Biogen, Maidenhead, UK



Component 1

Probability of re-identification due to leakage of identifiers:

$$\Pr(\text{reid}, \text{leak}) = \Pr(\text{reid} | \text{leak}) \times \Pr(\text{leak})$$

Where...

- $\Pr(\text{reid} | \text{leak}) = 1$ for direct identifiers
- $\Pr(\text{reid} | \text{leak})$ computed for quasi-identifiers based on empirically supported heuristics assuming that leak in two quasi-identifiers would lead to re-identification of patient

$$\Pr(X \geq 2) \quad \text{for } X \sim B(n_q, 1 - (r_q)^m)$$

Including “Hiding in plain sight” factor:

$$\times \Pr(X \geq 2 \text{ if } r_q \geq 0.7, \text{ or } Y \geq 2 \text{ if } r_q < 0.7)$$

for

$$X \sim B(n_q, h(1 - (r_q)^m)), Y \sim B(n_q, (1 - (r_q)^m))$$

- $\Pr(\text{leak}) = 1 - \text{recall}$ where recall is proportion of true annotation by total annotation using process (tool alone, tool+manual, manual only)



Component 2

- EMA quotes that “probability of re-identification of a record in a data set is 1 divided by the frequency of trial participants with same category/value of a set of the quasi identifiers (group size).” For example:

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

- Structured IPD is available to the trial sponsor but not to researcher, who received De-ID clinical documents.

Risk due to Component 2 will be calculated using IPD data considering quasi-identifiers appearing directly in clinical documents

Note: Populations proportion consideration not included here.



Scenario 1 – Conservative approach

Risk assessed considering component 1 and component 2 jointly, if quasi-identifiers are replaced with De-identified values from IPD data, will be given by

$$\underbrace{[Pr(reid \mid catch) \times [1 - Pr(leak)]]}_{\text{Component 2}} + \underbrace{[Pr(reid \mid leak) \times Pr(leak)]}_{\text{Component 1}}$$

Component 1 will be calculated as suggested in previous slides and component 2 will be major part of risk.

Component 2 will be maximum of individual risk for all subjects.

For example, referring table 1

- if all subject appeared in clinical reports as in table 1 then risk due to component 2 will be 1.
- If only subject 101 and 102 appeared in clinical report then risk will be 0.5.

This approach does not consider quasi-identifiers appearance in clinical reports.



Scenario 2 – Exact approach

If process of de-identification of clinical documents is able to detect the **information around quasi-identifiers** for each subject appeared in clinical document then risk assessed for component 2 will be more **exact**.

For example, referring table 1, if only sex appears for subjects in clinical reports then risk will be as below:

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

If subject 101 and 102 appears in clinical report with age information then risk due to component 2 will be 0.2.



Continued

What if all quasi-identifiers does not appear for all subject? For example: as shown in table below, age does not appear for two subjects i.e. 103 and 104

USUBJID	SEX	AGE	Equivalence Class	Re-id Risk
CT1/101	M	26	A1(3)	0.33
CT1/102	F	25	B(2)	0.5
CT1/103	M		A(3)	0.33
CT1/104	F		C(1)	0.5
CT1/105	M	26	A1(3)	0.33

- Re-id risk for subjects 103 and 104 will be calculated considering SEX variable info only i.e. $1/\text{number of repeated values in current dataset}$.
- Reid risk for other subjects would need some value for missing age of two subjects:
 - 1) *Assuming missing value could be any value and counting these subjects in each equivalence class*
 - 2) *Assigning missing values using probability calculated on remaining data.*
- In above example, it was assumed that subject with missing age will be counted in equivalence classes where value of other variable is matched. 103 is counted in equivalence class of sub 101 and 105



Thoughts on Component 1

Component 1 of re-id risk can be eliminated from calculation based on following justification:

- Process is robust enough and will not miss to annotate any identifier.
- 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.



Derive Value From Excellence ...

Q&A
Thank you