QUALIANCE

Data Transparency
Important Considerations for Data De-Identification

PhUSE SDE, London, 22. May 2014

Jean-Marc Ferran

Consultant & Owner

Disclaimer

- This presentation is sponsor-independent and represents my views
- I have no stock in any pharmaceutical company

QUALIANCE

Agenda

- EMA Guidance (Draft – June 2013)
- Data Transparency Concept
- De-identification methodologies
- De-identification rules and implementations
- Consideration for processes
- Open questions

QUALIANCE

EMA Guidance – Policy 70 (Draft) – June 2013 (1)

- Three types of Data:
  - (O) - **Open Access**—will be proactively disclosed
    - Most of clinical dossier – Clinical Overview, Clinical Summaries, and CSRs
  - (C) - **Closed Access**—will be publicly available through a controlled access process
    - Patient level data in CSR line listings
  - (CCI) - **May contain commercially confidential information** - will not be disclosed through this process
    - Biopharm and PK CSRs

QUALIANCE

EMA Guidance – Policy 70 (Draft) – June 2013 (2)

- "**Respect for the boundaries of patients' informed consent**: Patients participate in clinical drug trials in **the hope that their data will support the developmen**t and assessment of a **particular medicine that is useful for the treatment** of their disease, and will benefit the **advancement of science and public health**. The Agency takes the view that any other use of patient data oversteps the boundaries of patients' informed consent, and shall not be enabled by the policy."

- "The Agency will put in place measures to ensure the best-possible protection of public health (and regulatory decisions) against **claims resulting from inappropriate analyses**."

- **About Investigators, Health-Care provider, Authors of documents on the sponsor side:**
  - "This section contains personal data, such as the list of investigators; individual investigators' names, addresses, appointments, qualifications and clinical duties; similar information of other persons carrying out observations of primary or other major efficacy variables, such as a nurse, physician's assistant, clinical psychologist, clinical pharmacist or house staff physician; the author(s) of the report, including the responsible biostatistician(s). **The Agency takes the view that these persons have a role and responsibility for public health in ensuring the integrity of trial data and protecting patients' welfare. In light of the overriding public interest, these personal data are considered exempt from PPD considerations.**"

Source [1]

QUALIANCE

EMA Guidance – Policy 70 (Draft) – June 2013 (3)

- EMA recommends **Hrynaszkiewicz article (Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers) to de-identify data** in addition to specific rules
  - "A recommended minimum standard for de-identifying data is described in ***Hrynaszkiewicz***. In some situations, this minimum standard may need **to be supplemented by additional de-identification methods** (e.g. statistical). The methods of de-identification should be such that adherence will preclude subject de-identification, even when applying **linkages with other data carriers (e.g. social media).**"
  - "Agency is concerned **that emerging technologies for data mining and database linkage** will increase the potential for **unlawful retroactive patient identification**"

- **Data Standards**
  - "In future, **CDISC shall be the required standard**, in with future guidance from the Agency. **No conversion of formats is recommended**, either by the marketing-authorisation holder or the Agency."

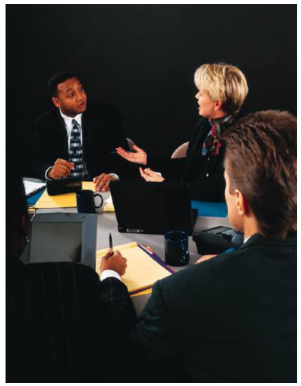
Source [1]

QUALIANCE

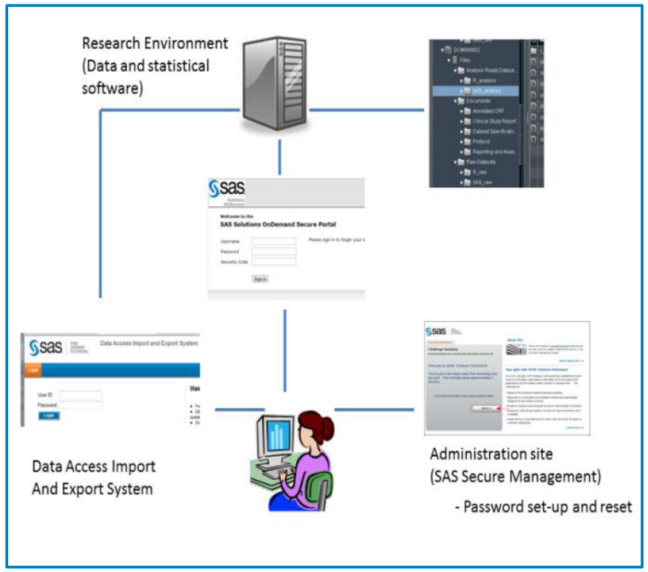
Data Transparency Concept (1)

Researcher writes research requests

Independent Review Board (IRB) evaluates requests



Secure Statistical Computing Environment



Data de-identification

QUALIANCE

Data Transparency Concept (2)
Data Sharing Agreement

- **Between Sponsor and Researchers**
  - Only **planed analyses** can be performed
  - Commitment to **publish analysis results**
  - Must **not intend to copy data**
  - Must **not intend to re-identify patients**
  - **Allow sponsor to use any invention** that comes out of the research and requires negotiation of any further rights in good faith
  - Researchers must **give access to their code**/tools
  - The research team must **include a statistician**
  - **Inform** regulatory authorities and sponsor of **any safety concerns** as soon as they are identified

QUALIANCE

Data Transparency Concept (3)
Informed Consent Form

- **Between Sponsor and Patients**
  - Patients give rights to use their data for the purpose of the study
  - In general Past informed consent forms do not address the Data Transparency concept directly
  - ***Companies assume that if the planned analyses are within the spirit of the protocol and anonymized data is provided, this is acceptable***
  - Companies also update their informed consents to be fully protected moving forward both with sharing data with external researchers or regulatory bodies doing it directly while privacy will be ensured.

QUALIANCE

Data Transparency Concept (4)
Data to be shared

- **Companies commit to consider requests based on data that are:**
  - Part of **successful submissions**
  - **Back in time** to a certain point: 2001, 2003, etc.
  - **Can be anonymized**
    - studies in rare Therapeutic Areas may not be eligible
  - Provide individual studies - **not pooled**
  - Involve an **Independent Review Board**
    - Systematically or only in case of rejection and want to run it through an IRB

QUALIANCE

Data protection directives

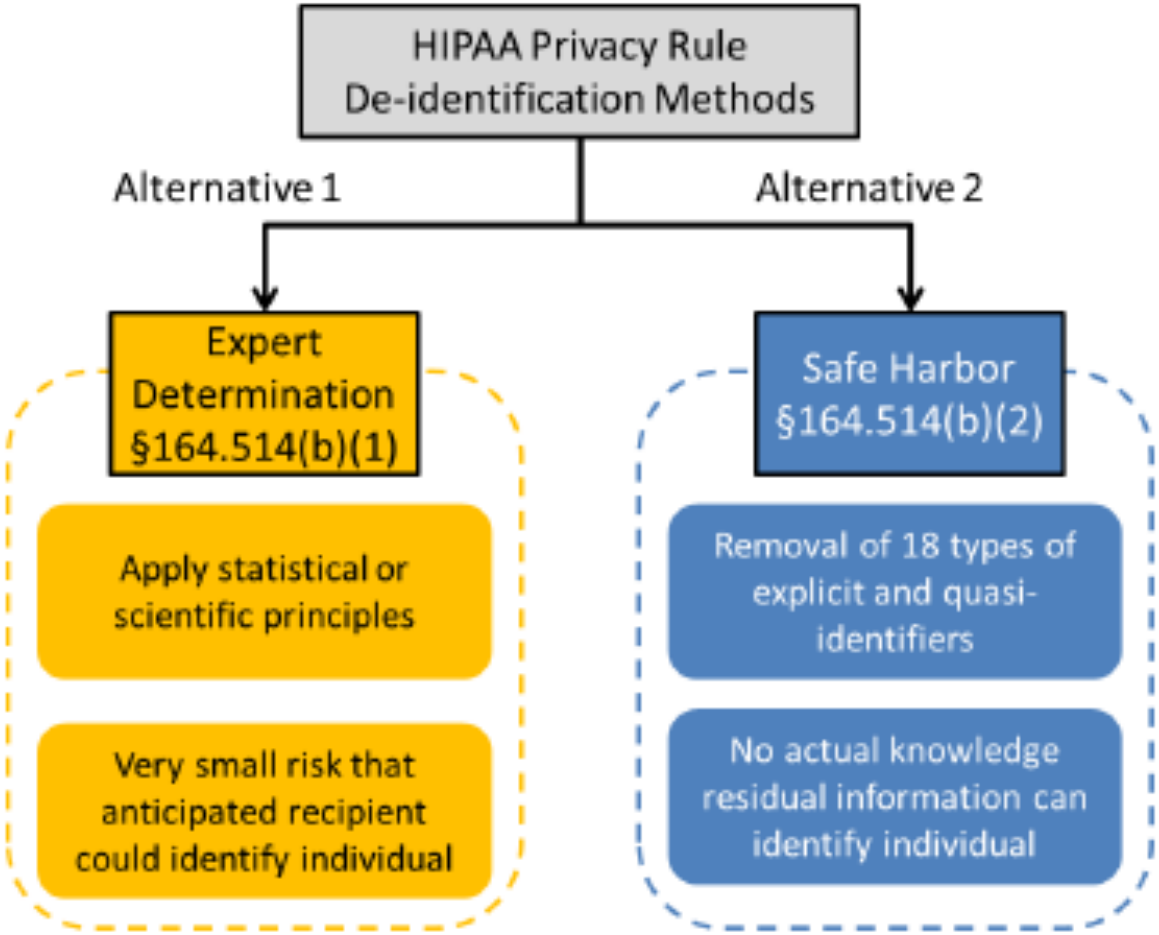
- **HIPAA**: Privacy Rule of the U.S. Health Insurance Portability and Accountability Act of 1996

- **EU Data Protection Directive** (do not provide specific technical guidelines)

- **Local European** member state directives

QUALIANCE

HIPAA de-identification methods (1)
Safe Harbor vs. Expert Determination



Source [2]

QUALIANCE

HIPAA de-identification methods (2) Safe Harbor (No residual information)

- Safe Harbor consists of removal of:

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(M) device identifiers and serial numbers
(E) Fax numbers	(N) Web Universal Resource Locators (URLs)
(F) Email addresses	(O) Internet Protocol (IP) addresses
(G) Social security numbers	(P) Biometric identifiers, including finger and voice prints
(H) Medical record numbers	
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique, identifying number, characteristic, or code
(K) Certificate / license numbers	
(L) vehicle identifiers and serial numbers, including license plate numbers	

Source [2]

QUALIANCE

HIPAA de-identification methods (3) Expert Determination (very small risk)

- **Expert determination**
  - **A person with appropriate knowledge of and experience** with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
    - (i) Applying such principles and methods, **determines that the risk is very small** that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
    - (ii) **Documents the methods and results** of the analysis that justify such determination;
  - It is important to recognize that while the re-identification provision does not permit assignment of a code or other means of record identification **that is derived from identifying individual information**, a covered entity may disclose such derived information **if an expert determines that the data meets the de-identification requirements** .

--> recoding of subject ids, offsetting of dates, etc.

Source [2]

QUALIANCE

Hrynaszkiewicz I et al. (1) Direct Identifiers

Name ([8-15])

Initials ([13])

Address, including full or partial postal code ([8-15])

Telephone or fax numbers or contact information ([8 10 12 15])

Electronic mail addresses ([8])

Unique identifying numbers ([8-15])

Vehicle identifiers ([8])

Medical device identifiers ([8])

Web or internet protocol addresses ([8])

Biometric data ([8])

Facial photograph or comparable image ([8 10 11 13])

Audiotapes ([11])

Names of relatives ([10])

Dates related to an individual (including date of birth) ([8 9 11 15])

Generalised HIPAA items 7-10, 18

Source [3]

QUALIANCE

Hrynaszkiewicz I et al. (2) Indirect Identifiers

Place of treatment or health professional responsible for care ([10 15])	Could be inferred from investigator affiliations
Sex ([9])	
Rare disease or treatment ([10])	
Sensitive data, such as illicit drug use or "risky behaviour" ([15])	
Place of birth ([10 15])	
Socioeconomic data, such as occupation or place of work, income, or education ([9 10 12 15])	MRC requirement is for "rare" occupations only
Household and family composition ([15])	
Anthropometry measures ([15])	
Multiple pregnancies ([15])	
Ethnicity ([9])	
Small denominators—population size of <100 ([14])	
Very small numerators—event counts of <3 ([14])	
Year of birth or age (this article)	Age is potentially identifying if the recruitment period is short and is fully described
Verbatim responses or transcripts ([15])	

Source [3]

--> If more than 3 in same study data, assessment by an independent researcher or ethics committee to evaluate the risk

QUALIANCE

Risk when combining identifiers

- ***Low**:* It has been estimated that the combination of ***Year of Birth, Gender,* and *3-Digit ZIP Code* is unique for approximately 0.04% of residents in the United States** [Sweeney 2007]. This means that very few residents could be identified through this combination of data alone.
- ***High**:* It has been estimated that the combination of a patient's ***Date of Birth, Gender,* and *5-Digit ZIP CODE* is unique for over 50% of residents in the United States** [Golle 2006, Sweeney 2002]. This means that over half of US residents could be uniquely described just with these three data elements.

Source [2]

QUALIANCE

De-identification rules (1)
Example

Table II. An example using fictitious data to illustrate the removal of personally identifiable information.

Centre ID	Investigator ID (INVID)	Investigator name (INVNAME)	Subject ID (SUBID)	Unique subject ID (USUBID)	Age (yrs)	AE start date	AE end date	Verbatim term
00123	279344	Dr Smith	5	TJF4392.005	57	29DEC2010	27JAN2011	Headache
00123	279344	Dr Smith	2	TJF4392.002	72	10JAN2011	06APR2011	Nausea
00123	279344	Dr Smith	1	TJF4392.001	91	25MAR2011	12AUG2011	Cold
00123	279344	Dr Smith	66	TJF4392.066	89	28MAR2011	31MAR2011	Cold
00123	279344	Dr Smith	8	TJF4392.008	94	01MAR2011	15MAY2011	Flu
05678	333721	Dr Jones	19	TJF4392.019	85	14OCT2010	20OCT2011	Cold
05678	333721	Dr Jones	4	TJF4392.004	53	24MAY2011	.	Headache
05678	333721	Dr Jones	23	TJF4392.002	76	01MAR2011	15MAR2011	Pain

New INVID | Remove INVNAME | New SUBID | New USUBID | Remove ages above 89 | Create age category | Add dummy dates | Add dummy dates | Remove

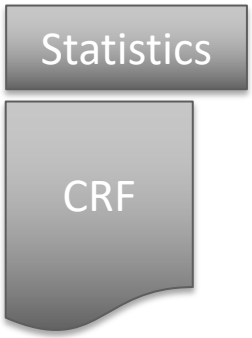
Centre ID	Investigator ID (INVID)	Investigator name	Subject ID (SUBID)	Unique subject ID (USUSID)	Age (yrs)	Age Category	AE start date	AE end date	Verbatim term
00123	227		8754	TJF4392.8754	57	<=89	19AUG2010	17SEP2010	
00123	227		5681	TJF4392.5681	72	<=89	06JUL2010	30SEP2010	
00123	227		1475	TJF4392.1475	.	>89	05SEP2010	23JAN2011	
00123	227		6589	TJF4392.6589	89	<=89	06SEP2010	09SEP2010	
00123	227		3562	TJF4392.3562	.	>89	29JUN2011	12SEP2011	
05678	208		1457	TJF4392.1457	85	<=89	16JUL2011	12SEP2011	
05678	208		2214	TJF4392.2214	53	<=89	04NOV2010	.	
05678	208		2236	TJF4392.2236	76	<=89	01JUL2010	15JUL2010	

Source [4]

QUALIANCE

De-identification rules (2)

- Recoding of
  - subject ids (and global ones)
  - investigator ids
  - lab ids
  - site ids
- Blanking of all verbatims/free text field
- Blanking of names
- Offsetting of dates
- Derivation of DoB into age and age category
  - For ex., patients over 89 years old grouped in category “>= 90”
- Grouping of sites and countries of less than 10 subjects
- Raw-level de-identification of
  - socio-economic information
  - Use of illicit drugs information or risky behaviour
  - Sensitive information (e.g., HIV, mental disease, venereal disease)
  - rare events



- Other in case of de-identification feasibility issues:
  - Grouping of countries to continents
  - Removal of all demographics

QUALIANCE

Consideration for processes

- **Deletion of the key / Use of restricted areas**
  - Deletion of any recoding tables or tables holding seeds
  - Deletion of QC outputs
  - Deletion of program logs
  - Capture in audit-trail

→ **Claim anonymization**

- **License issues** when sharing data? Ex: **medDRA**

- Provide **full or partial database** that fit the needs of the request?

- **Migrate data** to another data standards (e.g. CDISC)?

- Provide **raw and/or analysis** datasets?

- **Reassess de-identification** every 5/10 years?

QUALIANCE

Implementation

- **Challenge**: Different data models to consider for legacy studies
- **Implementation Phases**
  - Analysis -> De-identification plan
  - Implementation
  - Validation - Data is GxP until it is anonymized
  - Deletion of the keys (recoding tables)
  - Copy to Data Sharing Platform
- **Technical Implementation**
  - Macros
  - Scripts specific to data models
  - Metadata driven to address all data models

QUALIANCE

What documents to provide with the data?

- **Protocols** with any amendments (redacted)
- Annotated **CRF** (redacted)
- **SAP** (redacted)
- **Dataset specifications** (redacted)
  - Define.xml if available
  - **With documentation of de-identification**
- Redacted **CSR**
  - Appendices which include patient level data are not included
  - **Interaction needed between Biometrics and Clinical Reporting** to align data de-identification and documents redaction

QUALIANCE

Open questions

- What will **EMA decide** and how will this influence FDA and Japan?
  - And could that potentially make sponsors delay drug application in EU as a result?
- Will it be a **burden for Biotechs** for fund-raising if this becomes compulsory?
  - Funds used in something that may be seen as not directly adding value
  - Risk of loss of IP
- Will the opportunity be taken by **academics**?
  - Challenging results and methodology?
  - Or coming up with innovative hypotheses that may be confirmed through their research?
- **Rare Therapeutic Areas** seem to be at risk while they would benefit most?
- **Data standardization** will be key to success for data pooling across sponsors and so will De-Identification standards? One approach should emerge after some time.

QUALIANCE

References

- [1] **EMA Guidelines 0070 (Draft) – June 2013**
  - http://www.ema.europa.eu/docs/en_GB/document_library/Other/2013/06/WC500144730.pdf
- [2] **A De-identification Strategy Used for Sharing One Data Provider's Oncology Trials Data through the Project Data Sphere Repository**, Malin, 2013
  - https://www.projectdatasphere.org/projectdatasphere/html/resources/PDF/DEIDENTIFICATION&sa=U&ei=D69iU47DBYaN4ASf4YGgBw&ved=0CBsQFjAA&usg=AFQjCNGaWTa9-cXwUpP9q6UfE9FBjPV4vw
- [3] **Hrynaszkiewicz** I, Norton M L, et al. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. British Medical Journal 2010; 340:304–307
  - http://www.bmj.com/content/340/bmj.c181
- [4] **Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach** – Pharmaceutical Statistics 2014

QUALIANCE

Thanks!

Jean-Marc Ferran

Consultant & Owner, Qualiance ApS

dk.linkedin.com/in/jeanmarcferran/

@QualianceTwitta

QUALIANCE