Terminology Harmonisation in Data Sharing and Disclosure Terms and Definitions (version 2)

Term	Definition	Resource
Adversary	A data user who intentionally or inadvertently learns or discloses information about a data subject through re-identification or attribution. This user may be motivated by a wish to discredit or otherwise harm the organisation disseminating the data, to gain notoriety or publicity, or to gain profitable knowledge about particular data subjects. Data adversaries are sometimes referred to as intruders, snoopers or attackers	Definition adapted from Elliot, M., Mackey, E., O'Hara, K. et al. The Anonymisation Decision-Making Framework (2016). UK Anonymisation Network. Accessed at: https://eprints. soton.ac.uk/399692/1/The-Anonymisation-Decision-makingFramework.pdf (last accessed 24 March 2021).
Anonymisation	The overall process of protecting the privacy of data subjects, including clinical study participants, and reducing the risk of re-identification by 1) modifying (e.g. suppressing, obscuring, aggregating, altering) identifiable information in structured data and documents, 2) assessing and controlling the residual risk of re-identification and 3) considering the context of the data release.	Definition adapted from PHUSE: Data Anonymisation and Risk Assessment Automation, Version 1.0. (9 June 2020). Accessed at: https://phuse.s3.eu-central-1. amazonaws.com/Deliverables/Data+Transparency/ Data+Anonymisation+and+Risk+Assessment+Automation.pdf (last accessed 18 March 2021).
Anonymised data and documents	Data and documents that have been produced as the output of an anonymisation process.	Definition adapted from International Organization for Standardization: ISO 25237:2017(en) Health informatics – Pseudonymization. (January 2017). Accessed at: https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en (last accessed 23 March 2021) International Organization for Standardization: ISO/IEC 29100:2011(en) Information technology — Security techniques — Privacy framework (December 2011). Accessed at: https://www.iso.org/standard/45123.html (last accessed 24 March 2021).
Confidential business information (CBI)	In respect of a person (individual or organisation) to whose business or affairs the information relates, means business information that is not publicly available, in respect of which the person has taken measures that are reasonable in the circumstances to ensure that it remains not publicly available, and that has actual or potential economic value to the person or their competitors because it is not publicly available and its disclosure would result in a material financial loss to the person or a material financial gain to their competitors. (In reference to clinical reports submitted to Health Canada, as defined in Canada's Section 2 of the Food and Drugs Act.)	Definition adapted from Health Canada: Public Release of Clinical Information, Version 1.0. (12 March 2019). Accessed at: https://www.canada.ca/en/health-canada/services/ drug-health-product-review-approval/profile-public-releaseclinical-information-guidance.html (last accessed 18 March 2021).
Commercially confidential information (CCI)	Any information contained in the clinical reports submitted to the European Medicines Agency (EMA) by the applicant/MAH which is not in the public domain or publicly available and where disclosure may undermine the legitimate economic interest of the applicant/MAH.	Definition directly from European Medicines Agency: External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use (Policy 0070), Version 1.4. (9 November 2018). Accessed at: https://www.ema.europa.eu/en/humanregulatory /marketing-authorisation/clinical-data-publication/ support-industry/external-guidance-implementation-europeanmedicines-agency-policy-publication-clinical-data (last accessed 18 March 2021).

Data subject	An identified or identifiable natural person to whom a particular piece of data relates.	Definition adapted from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0 (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws. com/Deliverables/Data+Transparency /Protection+of+Personal +Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021). Garfinkel, S. L. (October 2015). 'De-Identification of Personal Information'. Internal Report 8053. National Institute of Standards and Technology. Accessed at: http://dx.doi. org/10.6028/NIST.IR.8053 (last accessed 18 March 2021).Elliot, M., Mackey, E., O'Hara, K. et al. (2016). The Anonymisation Decision-Making Framework. UK Anonymisation Network. Accessed at: https://eprints.soton. ac.uk/399692/1/The-Anonymisation-Decision-making-Framework.pdf (last accessed 24 March 2021).International Association of Privacy Professionals: Glossary of Privacy Terms. Accessed at: https://iapp.org /resources/glossary (last accessed 22 March 2021).
De - identification	A general term for any process of removing the association between a set of identifying data and a data subject present in data or documents. The association between data and subject is removed by modifying (e.g. removing, obscuring, aggregating, altering) identifiable information in structured data and documents.	Definition adapted from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0. (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency /Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021).Garfinkel, S. L. (October 2015). 'De-Identification of Personal Information'. Internal Report 8053. National Institute of Standards and Technology. Accessed at: http://dx.doi.org/10.6028/NIST.IR.8053 (last accessed 18 March 2021).Clinical Data Interchange Standards Consortium: Glossary, V15.0. (18 December 2020). Accessed at: https://www.cdisc.org /standards/glossary (last accessed 24 March 2021).
De-identified data and documents	Data and documents that have been produced as the output of a de-identification process.	
Direct identifier	Data that can be used to uniquely identify an individual (e.g. study participant ID, social security number, exact address, telephone number, email address, government-assigned identifier) without additional information or cross-linking other information that is in the public domain.	Definition directly from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0. (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency /Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021).
Equivalence class	Records (i.e. rows in a dataset) that share the same values for variables on a set of quasi identifiers.	Definition adapted from Information and Privacy Commissioner of Ontario: De-identification Guidelines for Structured Data. (June 2016). Accessed at: https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf (last accessed 18 March 2021). PHUSE: De-Identification Standard for CDISC SDTM 3.2, Version 1.01. (20 May 2015). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables /Data+Transparency/De-identification+Standard+for+SDTM+3.2+Version+1.0.xls (last accessed 22 March 2021). El Emam, K. (2013). *Guide to the De-Identification of Personal Health Information.* Auerbach Publications.
Individual patient or participant data (IPD)	The person-specific data separately recorded for each data subject in a clinical study.	Definition directly from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0. (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency /Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021).
Journalist risk	The risk of an adversary (individual or organisation) intentionally attempting to identify a data subject within a dataset. The adversary does not know if a specific individual is in the dataset.	Definition adapted from El Emam, K., & Arbuckle L. (2013). *Anonymizing Health Data.* O'Reilly
k-anonymity	A criterion used to ensure that there are at least k records within each equivalence class in a dataset.	Definition adapted from Elliot, M., Mackey, E., O'Hara, K. et al. (2016). *The Anonymisation Decision-Making Framework.* UK Anonymisation Network Publications. Accessed at: https://eprints.soton.ac.uk/399692/1/The-Anonymisation-Decision-making-Framework.pdf (last accessed 24 March 2021).
i - diversity	A refinement to the k-anonymity approach which assures that groups of records specified by the same identifiers have sufficient diversity to prevent inferential disclosure.	Definition directly from Garfinkel, S. L. (October 2015). 'De-Identification of Personal Information.' Internal Report 8053. *National Institute of Standards and Technology.* https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

Personal information (PI)	Subject-level data that can be linked to a data subject directly or indirectly, in particular by reference to details such as name, identification number, location data or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that subject.	Definition adapted from PHUSE: Data Anonymisation and Risk Assessment Automation, Version 1.0. (9 June 2020). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+Anonymisation+and+Risk+Assessment+Automation.pdf (last accessed 18 March 2021).
Privacy enhancing technology (PET)	Technologies that are designed to support privacy and data protection.	Definition directly from European Union Agency for Cybersecurity (ENISA): Privacy enhancing technologies (website). Accessed at: https://www.enisa.europa.eu/topics/data-protection/privacy-enhancing-technologies (last accessed 15 July 2021).
Prosecutor risk	The risk of an adversary (individual or organisation) intentionally attempting to identify a data subject within a dataset. The adversary does know that a specific individual is in the dataset.	Definition adapted from El Emam, K., & Arbuckle, L. (2013). *Anonymizing Health Data*. O'Reilly.
Protected personal data (PPD)	Any information relating to an identified or identifiable data subject; an identifiable subject is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to their physical, physiological, mental, economic, cultural or social identity.	Definition adapted from Directive 95/46/EC (Data Protection Directive) (24 October 1995). *European Union*. Accessed at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046 (last accessed 23 March 2021).
Pseudonymisation	A type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms. Typically, pseudonymisation is implemented by replacing direct identifiers (e.g. a name, a subject ID) with a randomly generated value.	Definition adapted from International Organization for Standardization: ISO 25237:2017(en) Health informatics – Pseudonymization. (January 2017). Accessed at: https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en (last accessed 23 March 2021). Clinical Data Interchange Standards Consortium: Glossary, V15.0. (18 December 2020). Accessed at: https://www.cdisc.org/standards/glossary (last accessed 24 March 2021). Garfinkel, S. L. (October 2015). 'De-Identification of Personal Information'. Internal Report 8053. *National Institute of Standards and Technology.* Accessed at: http://dx.doi.org/10.6028/NIST.IR.8053 (last accessed 18 March 2021). National Institute of Standards and Technology: Computer Security Resource Center Glossary. Accessed at: https://csrc.nist.gov/Glossary (last accessed 22 March 2021).
Pseudonymised data and documents	Data and documents that have been produced as the output of a pseudonymisation process.	
Quasi identifier	Data that in connection with other information can be used to identify an individual with high probability, e.g. age at baseline, race, gender, medical information, events, specific findings, location.	Definition adapted from Definition adapted from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0 (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021). PHUSE: A Global View of the Clinical Transparency Landscape – Best Practices Guide, Version 1.0. (22 May 2020). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Clinical+Trials+Data+Transparency+Toolkit+Best+Practices+Guide.pdf (last accessed 18 March 2021). PHUSE: De-Identification Standard for CDISC SDTM 3.2, Version 1.01. (20 May 2015). Accessed at:https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+Standard+for+SDTM+3.2+Version+1.0.xls (last accessed 22 March 2021). El Emam, K. (2013). *Guide to the De-Identification of Personal Health Information*. Auerbach Publications. PHUSE: Data Anonymisation and Risk Assessment Automation, Version 1.0. (9 June 2020). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+Anonymisation+and+Risk+Assessment+Automation.pdf (last accessed 18 March 2021).

Re-identification	Re-establishment of the association between a set of identifying data and the data subject found in data or documents.	Definition adapted from Garfinkel S, L. (October 2015). 'De-Identification of Personal Information'. Internal Report 8053. *National Institute of Standards and Technology*. Accessed at: http://dx.doi.org/10.6028/NIST.IR.8053 (last accessed 18 March 2021). De-identification Guidelines for Structured Data (June 2016). *Information and Privacy Commissioner of Ontario.* Accessed at: https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf (last accessed 18 March 2021). Computer Security Resource Center Glossary. National Institute of Standards and Technology. Accessed at: https://csrc.nist.gov/Glossary (last accessed 22 March 2021). **Re-identification risk** The probability that re-identification could occur. Definition adapted from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0 (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021). Definition adapted from Garfinkel S, L. (October 2015). 'De-Identification of Personal Information'. Internal Report 8053. *National Institute of Standards and Technology*. Accessed at: http://dx.doi.org/10.6028/NIST.IR.8053 (last accessed 18 March 2021). External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use (Policy 0070), Version 1.4 (9 November 2018). *European Medicines Agency.* Accessed at: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data (last accessed 18 March 2021).
Reference Population	The group of individuals that represent the basis for assessing the risk of re-identification. This group could be represented by the study population or a larger group of individuals.	Definition adapted from Health Canada: Public Release of Clinical Information, Version 1.0 (12 March 2019). *Health Canada.* Accessed at: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html (last accessed 18 March 2021).
Residual Risk	The risk of re-identification that remains on data or documents that have been produced as the output of an anonymisation process.	Definition adapted from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0 (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021) External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use (Policy 0070), Version 1.4 (9 November 2018). *European Medicines Agency.* Accessed at: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data (last accessed 18 March 2021).
Risk threshold	The maximum amount of acceptable re-identification risk remaining in documents and data after an anonymisation process has been applied. The threshold value can be either quantitative or qualitative.	Definition adapted from El Emam, K., & Arbuckle, L. (2013). *Anonymizing Health Data.* O'Reilly.
Safe Harbor method	This method describes 18 types of identifiers that must be removed in order for the resultant datasets to be considered de-identified according to the US Health Insurance Portability and Accountability Act (HIPAA).	Definition adapted from Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach (April 2015). *TransCelerate.* Accessed at: https://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-Data-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies.pdf (last accessed 23 March 2021).
Secondary use	Uses and disclosures that are different from the purpose(s) for which the data were collected as described in a clinical trial protocol and informed consent form.	Definition adapted from ISO 25237:2017(en). Health informatics – Pseudonymization (January 2017). *International Organization for Standardization*. Accessed at: https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en (last accessed 23 March 2021).

Sensitive information	Any data that, in the event of re-identification, could be considered harmful for a data subject in terms of employability, reputation, insurability, self-esteem or stigma, or could result in loss of income. The perception of information as sensitive is subjective and examples include substance abuse, mental disorders and abortion.	Definition adapted from PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0. (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf (last accessed 18 March 2021).
Single out	To isolate some or all records that identify a data subject in the dataset by observing a set of characteristics known to uniquely describe that data subject.	Definition adapted from Article 29 Data Protection Working Party: Opinion 05/2014 on Anonymisation Techniques, WP216. (10 April 2014). Accessed at: https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf (last accessed 18 March 2021). ISO/IEC 20889:2018(en). Privacy enhancing data de-identification terminology and classification of techniques. (November 2018). *International Organization for Standardization.* Accessed at: https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en:term:3.32 (last accessed 24 March 2021).
Synthetic data	Data that have been generated from one or more population models and which are designed to be non-identifying.	Definition adapted from Elliot, M., Mackey, E., O'Hara. K. et al. (2016). The Anonymisation Decision-Making Framework. *UK Anonymisation Network.* Accessed at: https://eprints.soton.ac.uk/399692/1/The-Anonymisation-Decision-making-Framework.pdf (last accessed 24 March 2021).
t-closeness	An equivalence class where the distance between the distribution of a selected attribute in the class and the distribution of the attribute in the whole table is no more than the value t.	Definition adapted from Article 29 Data Protection Working Party: Opinion 05/2014 on Anonymisation Techniques, WP216. (10 April 2014). Accessed at: https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf (last accessed 18 March 2021).