4 December 2018
EMA/796532/2018

Data anonymisation - a key enabler for clinical data sharing

Workshop report

30 November – 1 December 2017, European Medicines Agency, London

Disclaimer:

This workshop report only summarises the discussions that took place between participants of the referenced event. Accordingly, it does not represent or demonstrate the opinion of the European Medicines Agency on the matters discussed in it, and it cannot be understood and interpreted as guidance provided by the Agency. The European Medicines Agency does not assume any liability on the content of this workshop report.

30 Churchill Place ● Canary Wharf ● London E14 5EU ● United Kingdom
Telephone +44 (0)20 3660 6000 **Facsimile** +44 (0)20 3660 5555
Send a question via our website www.ema.europa.eu/contact

An agency of the European Union

Table of contents

1. Executive Summary

Technological developments are now providing increasing possibilities to store, mine, and analyse heterogeneous data across multiple data sources and multiple regions, offering the potential to derive novel insights from healthcare data. To maximise the benefit from healthcare data, pooling and integration with other datasets is required in order to extend potential insights beyond those derivable from a single study. In this context, data sharing can be defined as the practice of making original health data available for secondary research purposes by other investigators; data may be shared in various formats and the process of data release can range from sharing under open access arrangements to sharing under controlled and restricted conditions. Data must be shared in such a way as to ensure the protection of patient privacy. Whilst this is the foremost priority in any data sharing exercise, the changing technological and legislative landscape is increasingly challenging the ability to share data in sufficient depth and detail. This workshop explored how anonymisation of internationally sourced clinical trial data may be achieved while maintaining the scientific utility of the data in order to deliver benefits for the public good. While discussions extended to consider data generated through the delivery of normal clinical care and stored in patient registries, sharing of aggregated datasets as part of publishing research studies in a peer review journal publication was not considered.

Proactive sharing of clinical trial data has long been a key strategic aim of the European Medicines Agency (EMA) culminating in 2014 with the publication of Policy 0070 which led to the creation of a dedicated portal in 2016 that now hosts over 6650[1] documents. Data made publicly available via the portal are required to be fully anonymised before being hosted on the site. This approach while maximising access has resulted in many cases in substantial data redaction to ensure privacy thus reducing data utility. Further challenges hover on the horizon including the need to share more complex real world data including digitally captured data from wearables and smart devices.

A key objective of the meeting was to explore challenges around international data sharing since different legislative requirements governing data sharing result in differences between the US and EU. In both jurisdictions, anonymised data falls outside the scope of privacy legislation but differences in approaches to consent and re-consent, pseudonymisation, the handling of data from deceased persons, the use of data for secondary research including historical data and international data transfer challenge our ability to share not only multiregional clinical trial data but also digitally captured data from other sources.

Given these complexities, data anonymisation offers a route for clinical trial data sharing. However, the challenge is to determine what level of risk of re-identification is acceptable in order to deliver the potential benefits of data sharing. Assessment of risk is contextual: it requires a consideration of the circumstances and environment of release including the restrictions placed upon the data sharing, the sensitivity of the data, the potential linkage of released data with other data, and a continual re-examination of the robustness of anonymisation in the face of a changing data environment. Mechanisms are therefore needed to quantify the risk of re-identification, in order to understand the impact of different options for data release across the full range of healthcare data. The meeting explored the benefits and challenges of various potential options, including new innovative possibilities such as transformation, synthetic data approaches, encryption, and secure multiparty computation. Undoubtedly, utilisation of such approaches and the future development of additional novel approaches will be required, given the increasing interdependencies and connections of the data ecosystem,

[1] Data correct as of 7 September 2018.

evolving data analytics, machine learning, and artificial intelligence, all of which are causing a sea change in how data is interrogated and, therefore, data utility.

Over the last 5 years, increasing recognition of the value of data sharing has spawned a number of data-sharing platforms to enable access to individual patient level data; however, none of these is interoperable or integrated and measures for safeguarding the data vary across platforms. Moreover, given the fact that many trials are multi-regional, the technology and legal infrastructure associated with a platform should have global reach. In an evolving data landscape, platforms must anticipate supporting the sharing and integration of more 'challenging' data such as imaging, genomics, and real-world data. Currently most platforms consider that the responsibility for data anonymisation rests with the data controllers who contribute the data and not with the platform. This approach, however, may warrant re-examination, given that linkage of multiple datasets available on any platform(s) may increase the risk of re-identification.

Initial evidence suggested that utilisation of clinical trial data available through these platforms was disappointingly low, with potential reasons being the challenges of analysing data behind a firewall which may be inadequate for many questions, time restraints, lack of funding, lack of statistical support, requirements imposed by the platform for data access and the format the data is presented in e.g. non-parseable PDFs. Additionally, the lack of global anonymisation standards coupled with a lack of transparency around the chosen approach impacts negatively on the efficiency of data integration. While more recent information suggests demand is increasing, in order to enable the delivery of potential insights, data-sharing platforms should deliver an environment that simplifies the administrative process as much as possible. Importantly, data sharing should be valued and recognised in a similar way as peer reviewed publications, and appropriate metrics need to be developed to assign recognition to those generating and sharing the data[2].

Global guiding principles are urgently needed to address the needs of data anonymisation and utility and to find the appropriate balance to derive the benefits of data sharing. Meeting participants highlighted the need for international agreement on common terms describing the process of anonymising data and transparency on the choice of anonymisation approaches which, as far as possible, should be applicable globally to facilitate international clinical trial data sharing.

There was agreement that no data should, a priori, be exempt from data sharing but that a risk assessment should be performed based on a framework for anonymisation to determine the risk of re-identification and if, given the specific context, that risk were acceptable. This framework should offer methodologies to validate the risk assessment calculations and should incorporate the sensitivity of the data, the context of release, and the ultimate scientific utility of the data in its determination of the risk of re-identification. Furthermore, legislative requirements for the periodic evaluation of anonymisation in the light of the continually changing data environment necessitate associated metrics to allow monitoring of the process and assign accountability for maintaining an appropriate degree of anonymisation within the current data environment. Where data sharing may be limited for reasons of sensitivity or other concerns, other data sharing mechanisms such as strong data sharing agreements and controlled data access will be essential. Lastly, future innovation, research initiatives, and an ever evolving scientific landscape demand continuous and proactive exploration of new approaches to anonymisation.

[2] Such metrics should contribute to citation indexes such as the H index which contribute to tenure and agreed by funding bodies and academic institutions to encourage and facilitate data sharing among academics.

Importantly, where consent is used as the legal basis for data sharing, there must be clarity as to what will happen to the data, how it will be used, and how the rights of the data subject will be met at data disclosure and over time. Engagement with all stakeholders is necessary to communicate the benefits of data sharing but also to reassure stakeholders that the risk of re-identification is acceptably low. Ultimately, cultural change is needed to drive significant shift in behaviours coupled with mechanisms to enable not only safe data sharing but to provide a process that is legally compliant and reasonable for all stakeholders.

The sensitive and personal nature of healthcare data demands robust data protection combined with careful communication of the potential benefits of data sharing for scientific advancement and the public good. The benefits of data sharing should not come at the cost of privacy or personal autonomy. A global framework for anonymisation that (1) is able to meet the varied global legislative requirements, (2) highlights the methodology employed, and (3) assesses the current and future risk of re-identification is essential to build patient and public trust and confidence, and accountability, of the process.

2. Introduction

Proactive sharing of clinical trial data has been a key strategic aim of EMA since 2012 culminating in 2014 with the publication of Policy 0070 (EMA/240810/2013). This policy seeks to make clinical reports[3], subject to the applicable terms of use, publicly available in a proactive way to enable public scrutiny, not only of the company's data but also of regulatory decisions, and to inform future research by preventing or reducing duplication of effort. The policy allowed for a stepwise implementation with phase 1 leading to the creation of an online platform (https://clinicaldata.ema.europa.eu/web/cdp/home) in 2016 that, as of September 2018, hosts over 6650 documents from 126 regulatory procedures. Phase 2 of the policy, which has not yet been implemented, asked EMA to review the most appropriate way to make individual patient data (IPD) available while complying with privacy and data protection laws.

Since 2012 the landscape of clinical trial data sharing has changed significantly; a number of sponsors and journals now require data sharing statements and moving forward, as of January 2019, will require data sharing plans at the time of registration of a clinical trial (Taichman et al, 2017). Furthermore there are now several clinical trial data sharing platforms available[4] which enable access to IPD and which together have increased the momentum for data sharing. However, in tandem with these developments, the scientific and the legislative landscape has also changed, and ways must be found to share data in sufficient depth and detail so as to maximise its scientific utility while fully meeting data protection obligations. The legal framework in the European Union (EU) is now provided by the General Data Protection Regulation (GDPR). While GDPR does not change the core data protection principles of the previous Directives, it introduces new safeguards in the light of technological advances which must be considered for clinical data sharing. In parallel, technological innovations have driven the creation of huge amounts of data in near real time from a number of sources; while not all are healthcare-related, they create a data environment which significantly affects our ability to maintain

[3] Clinical reports shall mean the clinical overviews (generally submitted in module 2.5) and clinical summaries (generally submitted in module 2.7) and the clinical study reports (generally submitted in module 5, 'CSR'), together with appendices to the CSRs no. 16.1.1 (protocol and protocol amendments), 16.1.2 (sample case report form) and 16.1.9 (documentation of statistical methods), all in all PDF format.

[4] https://www.clinicalstudydatarequest.com/;https://www.bms.com/researchers-and-partners/independent-research/data-sharing-request-process.html;http://yoda.yale.edu/; https://www.pfizer.com/science/clinical-trials/trial-data-and-results; https://projectdatasphere.org/projectdatasphere/html/home; http://www.leo-pharma.com/Home/Research-and-Development/Clinical-trial-disclosure/Access-to-patient-level-data.aspx#; https://vivli.org.

privacy as, through the linkage of various data sources, patterns may emerge that enable re-identification of an individual from data that were previously anonymised. Layered on top of these considerations is the fact that data generation may be single or multiregional and data sharing is likely to be global, and thus these activities must comply with regulations across multiple jurisdictions. Additionally, the increasing interest in pragmatic trials with data drawn from electronic medical records or registries brings new and different challenges for data sharing.

There are therefore multiple challenges associated with the sharing of clinical trial data of sufficient depth and detail to be meaningful but still meeting the obligations of data protection. It is critical to find the balance between innovation and data protection, and embed 'privacy by design' (Danezis et al, 2014) into the planning of clinical trials in order to deliver benefits for the public good.

This background provided the impetus for the current workshop organised in collaboration with the Multi-Regional Clinical Trials Center, Brigham and Women's Hospital and Harvard, Boston, USA, that had three key objectives:

- To propose guiding principles to enable international data sharing for the benefit of public health;
- To build on the platform of work by EMA, to review anonymisation approaches applicable to a broader set of data which ensure privacy protection and meet the standards required to maintain accessibility and the scientific utility of the data;
- To examine opportunities for harmonisation of international clinical data sharing, taking into consideration data protection legislation in the different jurisdictions.

In the big data era a wider variety of sources of data will increasingly be utilised in clinical trials; however, the meeting focused on clinical trial data and only real world data in the context of patient registries and individual cohort studies. While it is anticipated that principles proposed for clinical trials will be relevant for other types of data, the following data sources were considered out of scope for the meeting:

- Electronic medical records;
- Claims/administrative health records;
- Social media data;
- Mobile health data.

The discussions were intentionally limited to EU and US legislation primarily for reasons of time and focus, and it is intended to test the principles against other jurisdictions as a subsequent exercise to validate their generalisability.

3. The global landscape for clinical data sharing

- *To describe the global landscape and highlight challenges for international clinical data sharing.*
- *To describe the current EMA guidance and its key recommendations highlighting the successes and challenges encountered during the implementation of Phase 1 of Policy 0070.*
- *To understand the legislation impacting clinical data sharing across two jurisdictions, exploring differences and similarities.*
- *To propose guiding principles to enable international data sharing in the public interest.*

3.1. International clinical data sharing: the context

"Researchers have a duty to make publicly available the results of their research on human subjects and are accountable for the completeness and accuracy of their reports. …..Negative and inconclusive as well as positive results must be published or otherwise made publicly available."

World Medical Association - Declaration of Helsinki, 2013

Re-analysis of clinical trial data and examination of the original analysis can not only reveal fresh perspectives and lead to new discoveries but importantly can reduce duplicative trials. It can thus avoid exposing patients/participants to unnecessary risk while allowing replication of research to test and strengthen trial results. However, such advantages can only truly be realised if participant privacy is robustly protected, if there is full participation from all trialists including academia, research companies, government and others, and if data are structured in such a way that they are interoperable and can be aggregated. Over the last 5 years, increasing recognition of the value of data sharing, coupled with information technological developments, has spawned a number of data-sharing platforms; however, none of these is interoperable or integrated and there is still minimal participation from academics or biotech companies, potentially because of the burden of the data sharing process, the lack of recognition or academic credit for data contributors, and the fear that intellectual property associated with the data will be lost (Bierer et al, 2017). It is also important to emphasise that sharing of clinical data in no way removes the fundamental obligation to register a clinical trial through the relevant portals[5] or to report the results.

A key objective of the meeting was to explore the challenges of data sharing across the US and EU and the impact of the different legislative requirements on data sharing between these jurisdictions. For example in each jurisdiction the sharing of anonymised data falls outside the respective legislation thereby avoiding the need to obtain consent for data sharing. In Europe data protection legislation requires a re-examination of whether the risks of re-identification change over time; in the US guidance allows for this possibility and while it does not require it, in practice it is often followed. As information technologies evolve and data availability in the environment changes this will be increasingly challenging to confirm. Currently a popular method for preserving anonymity for clinical reports is via redaction of any information which may increase the risk of re-identification of the individual; however, while effective, excessive data redaction can greatly compromise the scientific utility of the data. Furthermore, even if the risk of re-identification of an anonymised dataset is calculated to be acceptably small at the time of data disclosure, potential linkage of that dataset to others at an unknown future point in time, renders any calculation of the risk difficult or potentially unknowable. Thus while participants are overwhelmingly willing to share their research data when strong protections are in place, concerns about re-identification from previously anonymised data are likely to increase in an era where triangulation of clinical trial data with other data sources may defeat anonymisation approaches (Mello et al, 2013). Thus as a threshold principle, it must be appreciated that if scientific utility is to be retained, the risk of re-identification is never zero. The conversation must therefore be: can the residual risk of re-identification be defined; how significant is that risk; is it acceptably low under the circumstances; and should additional measures, beyond anonymisation, be considered? Acceptability will always be context dependent, influenced by the nature of the medical condition, its prevalence, the sensitivity of the data, and the autonomy of the individual. This meeting focused on the fundamental tension between maximising transparency and data utility while protecting

[5] https://www.ema.europa.eu/human-regulatory/research-development/clinical-trials/clinical-trial-regulation; https://www.fda.gov/scienceresearch/specialtopics/runningclinicaltrials/fdasroleclinicaltrials.govinformation/default.htm

patient privacy through anonymisation techniques and organisational measures (e.g. controlled access, data use agreements).

Complexities of data anonymisation, concerns around its impact on data utility and possibilities for data aggregation forces the need to consider sharing of pseudonymised data, as opposed to anonymised data. Under these circumstances informed consent for data sharing could be an important basis for processing personal data for secondary analysis. It is important to highlight that this consent would be in addition to that obtained for participation in the clinical trial. However, there are uncertainties associated with consent that limit its pragmatic application: it may be difficult to frame a consent request that anticipates how the data might be used in the future; a participant may withhold consent (thus rendering the anonymised dataset incomplete); the characteristics of those consenting to sharing their data may be quite different to those withholding consent (thus rendering the anonymised dataset unrepresentative); and the participant's views on privacy may change over time.

These challenges emphasise the timely nature of this workshop exploring how anonymisation of internationally sourced clinical trial data can be achieved while maintaining the scientific utility of the data.

3.2. EMA guidance and Policy 0070

To be open about its practice, the European Medicines Agency (EMA) issued a 'policy on publication of clinical data for medicinal products for human use' (Policy 0070). The first phase of Policy 0070 entered into force on 1 January 2015 and encompasses clinical overviews, clinical summaries, and clinical study reports for all human medicines submitted to the agency to support a request for a marketing authorisation. Importantly, the policy applies to all clinical trial data regardless of the outcome of the regulatory submission. Phase 2 of the policy which has not been implemented, will review available options for sharing of individual patient level data (IPD).

Phase 1 of Policy 0070 required the creation of a dedicated portal[6] where data is available in an open access manner with minimal restrictions; data must therefore be fully anonymised before being hosted on the site. In order to support marketing authorisation holders (MAHs) and following consultation with the European Data Protection Supervisor (EDPS), EMA provided guidance on the anonymisation of clinical reports that highlighted the then available techniques and those considered most appropriate; the guidance does not mandate any specific methodology. Critically, the MAH's anonymisation report to the Agency should carefully document the anonymisation process and the rationale for choosing the method, together with the results of an analysis of the risk of re-identification. Perennial challenges include the anonymisation of data from small populations (e.g. rare diseases, paediatrics) and data elements such as genetic data that appear to defy anonymisation approaches (Gymrek et al, 2013); both may drive excessive data minimisation to protect privacy. The impact of the anonymisation process on data utility should be evaluated with the goal of maximising utility while abiding by the data protection rules.

A number of further developments can be anticipated: firstly implementation of the GDPR would benefit from guidance from the data protection authorities (DPAs) around anonymisation approaches in the context of clinical data sharing to avoid significant heterogeneity in interpretation of GDPR; secondly, real world data (RWD), which is more complex and heterogeneous than traditional RCT data, will present new challenges for anonymisation approaches if data utility is to be maintained; thirdly,

[6] https://clinicaldata.ema.europa.eu/web/cdp/home

consideration will need to be given as to whether calculation of the risk of re-identification should be required or elective; and, finally, the mechanism for the release of IPD and potentially RWD data in the future will need to be decided.

3.3. Data sharing in the European Union and the United States

In the EU, the 2016 GDPR provides a single framework across the EU for protecting citizens' data in order to accommodate technological developments and globalisation, both a challenge and an opportunity in a digital world; constitutionalise the fundamental right to data protection as laid out in the Treaty of Lisbon 2009[7]; and harmonise a previously fragmented legislative framework. Additionally, creating a consistent legislative environment across EU member states and abolishing most prior notifications and authorisations will reduce bureaucracy and administrative burden.

The GDPR is comprehensive and its scope is wide: it includes all processing of personal data by automated or manual means, excluding processing by a person for purely personal or household activities; it covers both data controllers and processers in the EU as well as those outside the EU who collect and process EU citizens' data. Hence processing any data that originate within the EU, including from patients participating in multi-regional clinical trials, falls within the scope of GDPR even if the data controller and the processing of that data are outside of the EU. Equally data originating outside of the EU, potentially from patients who are not EU citizens would still fall under GDPR if that data were processed within the EU.[8] GDPR does not change the definition of personal data from the previous Directives; as such, anonymous data or data that is rendered anonymous falls outside the scope of GDPR. Notably, personal data that has been pseudonymised — that is, data that can be attributed to an individual through the use of additional information — does fall under the scope of GDPR.

Ensuring that terms are defined and clearly understood is critical, particularly across different legislative jurisdictions. GDPR defines several terms including personal data[9] (Article 4(1)), pseudonymisation[10] and anonymisation[11] (Recital 26), and personal data principles (fair and lawful processing[12], purpose limitation, data minimisation[13], clarity on further processing, clarity on data retention, data accuracy[14], accountability[15] and right to be forgotten[16]). De-identification[17] however a term used in the US is not used in GDPR.

[7] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12007L%2FTXT

[8] Article 3 of GDPR states "This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not." Recital (14) of GDPR states "(14) the protection afforded by this Regulation should apply to natural persons, whatever their nationality or place of residence, in relation to the processing of their personal data."

[9] **Personal data** - Any information relating to an identified or identifiable natural person, referred to as "data subject" - an identifiable person is someone who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his or her physical, physiological, mental, economic, cultural or social identity.

[10] **Pseudonymisation** - processing of personal data so that the data can no longer be linked to a specific person unless additional information (which is kept separate) is used. Please also refer to the legal definition given by Article 4(5) of the GDPR.

[11] **Anonymisation** - removal of the association between the identifying dataset and the data subject. As a result of anonymisation, the data must be stripped of sufficient elements such that the data subject can no longer be identified.

[12] **Processing** - any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.

[13] **Data minimisation** - The principle of "data minimisation" means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose. They should also retain the data only for as long as is necessary to fulfil that purpose. In other words, data controllers should collect only the personal data they really need, and should keep it only for as long as they need it..

[14] **Accuracy** - personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay.

GDPR provides a legal basis for processing personal data based on explicit (not tacit or presumed) consent that must be relatively specific and associated with a number of conditions (freely given, clear recording of consent, ability to withdraw consent without affecting the contract, and consent must be clear especially when for other purposes than the original [e.g. profiling]). Importantly a broad consent covering all secondary research is not legally acceptable as it cannot be sufficiently specific to embrace all future uses.[18] GDPR also allows EU countries to introduce further conditions and limitations on the processing of genetic, biometric and health data (Articles 9(4)). In order to reduce heterogeneity in the application of GDPR between EU member states, GDPR provides a new and improved governance framework for data protection which should deliver DPAs with improved co-operation. To further aid consistency there is a new decision making process for cross border cases and the creation of the European Data Protection Board (EDPB) for guidance and dispute settlement.

GDPR provides specific rules for the use of data for research purposes, subject to clear requirements set out in Articles 89 and 9 (2)(j). Notably, such powers are in the context of appropriate safeguards for data subjects which include rights to transparent and clear language; rights of information; rights of access; right to object, correct, delete or block; right not to be subject to automated decision making; right to portability; and clear time limits around data holding and obligations for the data controller that are graduated depending on the nature and potential risks of the processing operations.

Particularly pertinent to the workshop were new rules addressing international transfer of data. GDPR provides clear rules for when EU law is applicable but also supports international transfers with a new and improved tool kit. If a body does not comply with GDPR, substantial sanctions can be applied, which can reach up to 4% of global turnover, depending on the nature, duration, and gravity of the infringement. Importantly, an entity can be held liable if the data it has released to another entity results in identifying an individual because of a change in the external data environment.

The US legislative environment is principally shaped by the US Health Insurance Portability and Accountability Act of 1996 (HIPAA); it governs the use and disclosure of protected health information (PHI) by 'covered entities' (defined as health care plans, clearinghouses, and providers who electronically transmit any health information in connection with certain transactions) including the conditions under which PHI may be used or disclosed by covered entities for research purposes[19] (HIPAA Privacy Rule). The Privacy Rule also confers rights on individuals, including rights to access or amend their PHI and to obtain a record of when and why it has been shared. Thus, healthcare professionals conducting clinical studies or participating in clinical trials may be affected indirectly or directly by the Privacy Rule depending on their relationship with covered entities upon whom they may rely, for example, to provide research support or be a source of PHI. Under the Privacy Rule, covered entities need a legal basis to disclose PHI to pharmaceutical companies and others participating in the clinical trial process who need personal health data of trial participants. This may take the form of an individual’s written permission for the specific research study, termed an authorisation, or a waiver of

[15] **Accountability** - Principle intended to ensure that controllers are more generally in control and in the position to ensure and demonstrate compliance with data protection principles in practice. Accountability requires that controllers put in place internal mechanisms and control systems that ensure compliance and provide evidence – such as audit reports – to demonstrate compliance to external stakeholders, including supervisory authorities.

[16] **Right to be forgotten** - the right of the data subject to the erasure of personal data concerning an individual without undue delay (on the grounds explained in Article 17 of the GDPR).

[17] **De-identification** - a term used in the US to describe the process used to prevent personal identifiers in a dataset from being connected with information so as to allow identification of an individual. De-identification does not describe a single technique but rather a collection of approaches (e.g. suppression, averaging, generalization, perturbation, swapping) that can be applied to data with different levels of effectiveness. De-identified data may still allow a data generator or a trusted party to retain the means (e.g. a code, algorithm or pseudonym) to identify the person.

[18] In the US, the revised Common Rule allows for broad consent as discussed further below (see page 14).

[19] https://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf

authorisation that may be obtained through an Institutional Review Board or Privacy Board.[20] Additionally, the Privacy Rule introduces flexibility through the ability to use and disclose PHI included in a limited dataset[21] without obtaining authorisation or waiver of authorisation via the use of a data use agreement between the covered entity and the dataset recipient. A covered entity may always use or disclose health information that has been de-identified for research purposes; most PHI is therefore currently delivered as de-identified information, usually with a unique coded identifier (maintained separately and not shared with the dataset recipient. In this case HIPAA no longer applies. A critical difference between HIPAA and the GDPR is the application of the respective regulations to coded information in which the code is retained: in the US, 'de-identified' data is no longer subject to HIPAA whereas, in the EU, if any code is retained, GDPR applies.

HIPAA provides for two methods of de-identification:

- Safe Harbor: Requires removal of 18 identifiers[22] specified in the regulation that relate to the individual, and the individual's relatives, employers or household members. Notably these 18 identifiers do not specifically include genetic data[23], potentially secondary to the fact that HIPAA is now more than 21 years old. HIPAA was amended in 2013 by the Omnibus Final Rule to address genetic information.[24]
- Expert Determination: A person with 'appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods' of de-identification determines that risk of re-identification of a dataset either alone or in combination with other information is 'very small'. While there is no strict definition of what would be 'appropriate expertise', the Office of Civil rights would review the relevant professional experience and academic or other training of the expert used by the covered entity, as well as actual experience of the expert using health information de-identification methodologies.

[20] https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/research/index.html. Accessed 5 July 2018.

[21] Refers to PHI that excludes 16 categories of direct identifiers and may be used or disclosed, for purposes of research, public health, or health care operations, without obtaining either an individual's Authorisation or a waiver or an alteration of Authorisation for its use and disclosure, with a data use agreement.

[22] (1) names, (2) 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: (i) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (ii) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000, (3) 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 elements of 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, (4) telephone numbers, (5) vehicle identifiers and serial numbers, including license plate numbers, (6) fax numbers, (7) device identifiers and serial numbers, (8) e mail addresses, (9) web universal resource locators (URLs), (10) social security numbers, (11) internet protocol (IP) addresses, (12) medical record numbers, (13) biometric identifiers, including finger and voice prints), (14) health plan beneficiary numbers, (15) full face photographs and any comparable images, (16) account numbers, (17) any other unique identifying number, characteristic, or code, except as permitted for re-identification purposes, (18) certificate/license numbers (extracted from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#protected (18 October 2018).

[23] Genetic information includes information about the genetic tests of an individual and his or her family members, the medical history of those family members, and any request for genetic services (including genetic testing, counseling, or education) or participation in clinical research. A family member includes any dependent or relation to the fourth degree (e.g. great- great-grandparents or grandchildren, children of first cousins) or closer, without reference to the existence of biological ties.

[24] Department of Health and Human Services, Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule, 78 Fed. Reg. 5566 (Jan. 25, 2013). The protections under HIPAA extend to all genetic information, whether the information originated before or after the compliance date for the final regulations.

In the US, research involving human subjects[25] that is conducted or supported by any Federal department or agency is subject to the Federal Policy for the Protection of Human Subjects or the 'Common Rule'[26]. Further, if human subjects[27] research now involves products regulated by the Food and Drug Administration (FDA) (e.g. food and color additives, drugs for human use, medical devices for human use, biological products for human use, and certain electronic products) whether or not conducted or supported by any Federal department or agency, it is subject to regulation by the FDA.[28] FDA's regulations on investigational drugs, biological substances and medical devices largely track the Common Rule. HIPAA aims to protect the privacy of a person's health information while the Common Rule and FDA regulations address basic ethical principles and processes to protect subjects participating in clinical research. A revised Common Rule was published in January 2017, effective January 2019, and FDA intends to harmonise its own regulations in line with this revised version. Important aspects of the revision in the context of the current discussion include the introduction of broad consent for unspecified future use of biospecimens and data; new required elements in the consent form; expansion and change in procedure for research that qualifies for exempt review where the primary risk of the research is breach of privacy and confidentiality. A further important distinction is that the Common Rule only applies to living subjects while HIPAA applies for 50 years after the death of an individual; in contrast, GDPR does not apply to deceased persons or their data (although EU countries may impose rules on the processing of personal data of deceased persons). Thus there are clear differences between the requirements of the Common Rule and HIPAA on the one hand and GDPR on the other, specifically around the inclusion of broad consent, the jurisdiction over coded but de-identified data, and the methodology of anonymisation. The complexities and differing applicability of HIPAA, the Common Rule, and the FDA regulations creates a complicated landscape. One advantage of GDPR is that it provides for one set of principles across datatypes, sources, and methods of collection.

GDPR is built upon the fundamental principle that personal data should only be processed for a stated purpose and, once that purpose is completed, the data should then be deleted or anonymised as the legal basis for holding the data is lost. While anonymisation may therefore enable the use of historically collected data for secondary research, the inability to link across different datasets for a given individual will significantly impact the scientific utility of the data. Further challenges include: international data transfer and sharing, given the global nature of research and the data environment; maintaining anonymisation in the light of technological advances not anticipated when the data were initially anonymised; how legislation in the two jurisdictions views anonymisation with its known non-zero risk of re-identification; how that risk of re-identification, both initially and over time, can be evaluated and quantified; how to define a 'very low probability' of re-identification; where responsibility for the auditing and policing of the process lies; what measures should be taken if the regular monitoring and re-assessment lead to the conclusion that the re-identification risk has changed

[25] A human subject means a living individual about whom an investigator (whether professional or student) conducting research obtains (1) Data through intervention or interaction with the individual, or (2) Identifiable private information. https://www.hhs.gov/ohrp/sites/default/files/ohrp/policy/cdebiol.pdf. Accessed 1 December 2018.

[26] https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html#46.102. Accessed 5 July 2018.

[27] Note that the definition of 'human subject' differs between the FDA and HHS regulations. FDA defines 'human subject' as "an individual who is or becomes a participant in research, either as a recipient of the test article or as a control. A subject may be either a healthy individual or a patient." For a comparison of the differences between these two sets of regulations see https://www.fda.gov/ScienceResearch/SpecialTopics/RunningClinicalTrials/EducationalMaterials/ucm112910.htmAccessed 1 December 2018.

[28] See for a complete listing of the regulations applicable to clinical trials: https://www.fda.gov/scienceresearch/specialtopics/runningclinicaltrials/ucm155713.htm. Accessed 5 July 2018.

and the impact of requests to remove the data (right to be forgotten) on secondary research and already released datasets.

4. The Foundations of Data Anonymisation

- *To define and critique the key concepts which must be considered from a technical (methodological) and legal perspective.*
- *To discuss the balance between data anonymisation and scientific utility and how it can be achieved.*
- *To consider how the context of the disease affects the risk-based approach.*
- *To explore any international differences.*

4.1. Functional anonymisation and the data environment

Anonymisation provides a mechanism to manage the tension between safeguarding personal privacy and maximising the utility of data. As discussed earlier, unless the data are completely random, anonymisation can never be absolute if any data utility is to be retained; it is important, therefore, to avoid success terms when describing anonymisation such as 'truly anonymised'. Rather the challenge is to understand 'an acceptable level of risk' and the determination of that requires consideration of the downstream impact of any potential confidentiality breach. The impact may be difficult to define, as it will depend not only on the data but also upon the context of release and the resultant data situations that arise from the use of the data and the data interacting with the data environment. Moreover, the data environment (Elliot and Mackey, 2014) is usually dynamic and complex depending on the constraints under which data are shared and released. While challenging to describe, in general the data environment may be described by reference to four parameters:

- Agents – generally considered to be people but increasingly may be artificial intelligence and machine learning systems;
- Other data within the local or global environment;
- Data governance which determines who can access the data and what restrictions on use are placed upon users;
- Security infrastructure.

Whether data are anonymised or not is a function of the relationship between those data and the specific data environment: the same dataset will have different risks under different circumstances. It may be argued, therefore, that it is impossible to determine risk by considering the data alone: a framework is needed for developing an anonymisation policy that conceptualises, frames, and clarifies each individual data situation. The UK anonymisation framework[29] provides a practical tool to think constructively about the individual data context with a ten-step process: to understand the level of risk, to manage that risk, and to enable the effective anonymisation of personal data. It gives guidance on the three key areas:

[29] http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf. Accessed 17 January 2018.

(i) Audit of the data situation (systematically describing the data and their environment; how the data are to be used and the legal and ethical context of the data);

(ii) Analysis and control of risk (assessing the risk of disclosure from the data situation and how the risk is to be managed), and

(iii) Impact management (what needs doing before sharing or releasing data so as to maintain a very low risk of re-identification and what should be done in the case of unintended re-identification or security breach).

While this framework has been developed in the setting of UK legislation, the general principles are broadly applicable and should enable us to move closer to a global harmonised concept for anonymisation.

4.2. Risk-based approaches for data anonymisation

Interaction between the data and the data environment or context determines the overall risk of identification. The challenge lies in the need to balance privacy with data utility and find the acceptable trade-off between these two competing needs.

Defining overall risk is complex and is a function not only of the data environment but also of the dataset itself and the elements it contains, coupled with the specifics of the disease. Thus while anonymisation may well be secure in a contained (e.g. not public or open) system, how it operates in the context of a continually changing data environment that may alter the identifiability of the original data must be understood. Existing standards and guidelines generally divide data elements into two groups, direct identifiers and quasi-identifiers[30], both of which need to be addressed during anonymisation since there have been multiple examples of successful re-identification attacks using quasi-identifiers (Robertson, 2013; Sweeney, 1997; Sweeney, 2013; El Emam et al, 2011) especially when combined with other publicly available information. Mechanisms are therefore needed to quantify the risk of re-identification, in order to understand the impact of different options for data release across the full range of healthcare data. The process must consider the risks posed by the dataset itself, e.g. number of quasi-identifiers, the data to population or population to data risk[31] (El Emam, 2012, Danker et al, 2012), the context of release and the layers of protection in place including the extent to which it has been perturbed (or modified), security and privacy controls, and contractual controls. Data utility can be severely affected as a result of using an overly conservative reference population. For example, for public data release, such as that employed for Phase I of Policy 0070, there is no context to evaluate: it is assumed that an adversary exists, has public access (and can download the data and link it to other existing datasets), and will attempt to re-identify individuals. Therefore, no external controls can be put in place to manage the risk beyond modifying the data. While delivering maximum access, such an approach can result in excessive caution e.g. excessive redaction or removal of all quasi-identifiers in the anonymisation approach.

Practical but difficult questions that must be answered are: when are data modifications sufficient to achieve anonymisation? What threshold of re-identification risk is acceptable? How does GDPR consider re-identification that may occur following linkage to other datasets, whether anticipated or not?

[30] **Quasi-identifiers** - pieces of information representing a person's background information (e.g. their date of birth, clinic visit, residence postal code, sex and ethnicity) that are not of themselves unique identifiers but which can be combined with other quasi-identifiers and become personally identifying information.

[31] The data to population risk is determined by the uniqueness of the individual, that is the number of individuals in the reference population (for example the clinical trial, similar clinical trials, people with the illness, etc.) and which share the same characteristics, and hence quasi-identifiers, of an individual in the dataset to be released. Thus an adversary may choose to start the attack by targeting an individual either in the dataset or in the population.

Answering these interrelated questions involve selecting an appropriate risk metric, a suitable threshold and the actual measurement of the risk in the clinical information to be disclosed.[32] The choice of metric depends on the context of release but for public release, with no controls a maximum risk of 1 or 100% to attempt re-identification is assumed. Based on the recommendations made in the Institute of Medicines report[33] and the available precedents for public release of health data, EMA has previously stated that it is advisable to set the threshold to a conservative level of 0.09[34]. However, different organisations choose different thresholds depending on the characteristics of the data source, e.g. size and nature of the population of disease studies, number of direct and quasi-identifiers in the data, number of participants in the study, number and geographical distribution of study centres.[35] Recital 26 states that "account should be taken of all means reasonably likely to be used (…) to identify the natural person either directly or indirectly" and provides examples of what 'reasonable' may mean, but there remains scope for interpretation[36]. However, thresholds of risk of re-identification of less than 5% or 0.05 will result in a marked loss of data utility. Furthermore, in the era of big data with potentially multiple data linkage opportunities and new analytical methods, e.g. data mining, a larger number of attributes will need to be considered in the assessment of risk. Knowing when to stop modifying data to improve anonymisation is key in order to achieve the appropriate balance between privacy protection and data utility.

4.3. How does consent influence the data anonymisation approach and data sharing across different regulatory jurisdictions?

As a general principle, data should be processed in compliance with the spirit of the data protection legislation. Article 6 of the GDPR provides for six legal grounds for lawful processing of personal data: consent, contractual necessity, compliance with a legal obligation, the protection of the vital interests of the data subject or of another natural person, public interest, and legitimate interests pursued by the controller or by a third party. As discussed earlier if data is anonymised (so that it is no longer related to an identifiable person) then it is no longer considered as personal data and its processing falls outside the scope of GDPR.

Where consent (defined by Article 4 (11) of the GDPR[37]) is used as a legal basis for processing personal data, Articles 6(1)(a), Article 9(s)(a) and Article 7 contain relevant provision e.g. that it should be 'freely given, specific and informed' . Nevertheless, Recital 33 acknowledges that it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection and gives further guidance in such cases.

[32] External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use accessible at https://www.ema.europa.eu/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-1.pdf

[33] Sharing Clinical Trial Data: Maximising Benefits, Minimising Risks,; https://www.nap.edu/download/18998#

[34] https://www.ema.europa.eu/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-1.pdf

[35] Evaluation of re-identification risk for anonymised clinical documents. PhUSE, 2017, Paper RG02; https://www.phusewiki.org/docs/Conference%202017%20RG%20Papers/RG02.pdf

[36] To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.

[37] **Consent** - consent refers to any freely given, specific and informed indication of the wishes of a data subject, by which he/she agrees to personal data relating to him/her being processed (see Article 2 sub (h) of Data Protection Directive 95/46/EC and Article 2 sub (h) of Regulation (EC) No 45/2001. Consent is an important element in data protection legislation, as it is one of the conditions that can legitimise processing of personal data. If it is relied upon, the data subject must unambiguously have given his/ her consent to a specific processing operation, of which he/she shall have been properly informed. The obtained consent can only be used for the specific processing operation for which it was collected, and may in principle be withdrawn without retroactive effect.

Some points need further consideration including:

- respecting consent in secondary processing, data-sharing and data-intensive research (i.e. ensuring that the extent of the consent gained for the primary purpose is respected in the secondary processing of data either by the original researchers or when the data are shared with others);
- developing realistic ways to maintain and refresh consent over time (when consents are increasingly regarded as perishable) in biobanking, data-sharing, and data-intensive research;
- developing realistic ways to re-consent an individual when a child attains majority in the context of longitudinal studies;
- documenting, depositing, and storing evidence of informed consent such that it can be properly acted upon in the future.

We need to consider if and how informed consent forms can be written so as to inform, anticipate, and explain future possibilities. Harmonisation of practices compliant with the GDPR within Europe may be fostered by specific Codes of Conduct adopted under Article 40 (GDPR) that brings together EU thinking and emerging international perspectives on the governance of data sharing.

The GPDR requires that data be collected for specified, explicit, and legitimate purposes and must not be further processed in a way that is incompatible with those purposes (GDPR Article 5 (1(b)). Recital (50)[38] and Article 6(4) state that processing can be allowed for purposes other than that of which the personal data have been originally collected where it is compatible with the original purposes for collecting the data. Thus, if further processing of data is considered compatible with the initial purpose a separate legal basis from that which allowed the collection originally is not required. According to Recital (50) further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be considered to be compatible lawful processing operations, and these scenarios are governed by Article 89 of the GDPR as specific processing situations. For other cases, Article 6.4 sets out five tests to determine if processing for a further purpose is compatible with the purpose for which the personal data were initially collected:

- any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;
- the context in which the personal data have been collected, in particular the reasonable expectations of the data subjects based on their relationship with the controller;
- the nature of the personal data;
- the possible consequences of the intended further processing for data subjects;
- the existence of appropriate safeguards, which may include encryption or pseudonymisation.

It needs to be further investigated whether compatible processing or Article 89 may offer a credible lawful route for processing healthcare data for secondary research.

[38] The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. In such a case, no legal basis separate from that which allowed the collection of the personal data is required. If the processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, Union or Member State law may determine and specify the tasks and purposes for which the further processing should be regarded as compatible and lawful.

4.4. Defining sensitive data – influence of the context of the disease on the tolerability of risk

4.4.1. Rare Diseases

Patients, data, and biosamples are particularly scarce in rare diseases and hence the need for collaboration and data sharing is critical in order to improve understanding of the mechanisms underlying these diseases and to develop disease modulating treatments. However, rare diseases pose particular challenges around data sharing as it is difficult to anonymise data from small trials, from small patient populations or those with specific genetic mutations without significantly reducing data utility. A Delphi survey of 15 expert patients and their representatives by EURORDIS (the European Organisation for Rare Diseases) reported widespread agreement on the urgent need to share data in order to accelerate the development of treatments for rare diseases[39]. Importantly, the respondents favoured having a trustworthy stakeholder as the main curator of their data. However, at focus group meetings (held in 2014 during EURORDIS meetings), most participants felt very strongly that people should be contacted again if the specific use of their data were not covered by the original consent; a view which persisted regardless of the use of anonymisation or the extreme rarity of the disorder from which individuals suffered (McCormack et al, 2016). There is evidence that such sentiments are not unique to rare diseases with patients across a broad range of disease strongly supporting data sharing, largely irrespective of the purpose for which the data would be used, provided adequate safeguards were in place (Mello et al, 2018).[40]

Fears about re-identification may also be influenced by the personal or geographical proximity of the person who identifies the data subject. For example the person may be concerned if a neighbour or colleague becomes aware of their rare disease and yet may feel comfortable with posting personal information online because the digital risk is perceived as more distant or abstract.

From the perspective of an individual with a rare disease, important conditions for sharing data include: consideration of the data that need to be collected in advance of authorising a clinical trial; third parties, who should be identifiable and identified in the consent, should be qualified and able to explain their objectives and their results to the data collector (or EMA) before the results are placed in shared either in a controlled access environment or in the public domain; and individuals should consent to data sharing and their consent sought again when conditions change and when appropriate.

4.4.2. Mental Illness

GAMIAN-Europe, a Global Alliance of Mental Illness Advocacy Networks across Europe,[41] is concerned specifically with individuals with mental illness, for whom re-identification from personal information can lead to a number of adverse consequences including discrimination in employment, reduced access to insurance including healthcare, and inability to secure loan and credit advances. As such, surveys of patients with schizophrenia or depression, found that reluctance for data sharing stemmed from their previous experience of discrimination (Thornicroft et al, 2009) and over 70% of patients with schizophrenia and with depression wanted to conceal their diagnosis in data collected in any format

[39] Delphi exercice, first round, 06/10/2015 to 04/12/2015. RD Connect project 2012-2018 (http://rd-connect.eu). Expert patients and representative's views on an international platform for sharing data and bio-specimens.

[40] The one exception was that fewer participants were willing to share their data for use in litigation.

[41] Gamian-Europe (Global Alliance of Mental Illness Advocacy Network) represents the interests of persons affected by mental illness and advocates for their rights. Its main objectives are: advocacy, information and education, anti-stigma and discrimination, patients' rights, co-operation, partnerships and capacity building.

(Lasalvia et al, 2013). Furthermore while not unique for such illnesses, people's online activity (e.g. comments and 'likes' on Facebook or online shopping) can be aggregated and powerful algorithms used to predict personality traits and personal circumstances (Kosinski et al, 2013). Thus while digital advances are of value for the monitoring and treatment of mental illness, it must be appreciated that as for all patients, they can also compromise privacy, particularly when seemingly unrelated and/or innocuous data sources are combined.

4.4.3. Paediatric Patients

Paediatric clinical trials are often conducted across multiple national centres (Lepola et al, 2016) due to frequently small patient populations, the rarity of some of the conditions, and limited specialist facilities. In order to facilitate high quality ethical studies in children and increase the availability of medicines for this population, in 2011 EMA established the European Network of Paediatric Research (Enpr-EMA[42]). One of the vehicles to deliver on its mission is assisting and communicating with ethical committees on issues relevant to research and clinical trials. This is particularly important given the heterogeneity in the consent and assent[43] requirements for paediatric clinical trials with 3 different frameworks covering just the European Economic Area and different ages for independent consent (age of majority)[44]. eYPAGnet, a European Young Persons' Advisory Group network[45],which has the mission to improve the capacity of collaboration across the range of stakeholders involved in paediatric research and medicines development recently became a member of Enpr-EMA. eYPAGnet surveyed young people aged 12 to 23 years, including those with a significant health condition and some clinical trial participants. Of 11 young people responding, 9 said they had no concern about sharing their anonymised data in order to increase knowledge of their disease. However, only 4 out of 10 respondents who participated in a trial remembered being told what would happen with their data.

A number of concerns particularly affect young people. First and foremost as knowledge and understanding of a medical condition increases, young people may become more involved in discussions around their treatment with healthcare professionals. As a result, it is likely that an individual's understanding and opinion on which data are held and what they might be used for, will change with age. Thus the ability to give consent for data sharing depends upon a child's maturity and the development of adequate understanding about their rights to (or need for) privacy. Young people may also be at a higher risk of re-identification because of their engagement with social media and thus the risk of social and employment consequences of re-identification for young people may be increased.

[42] https://www.ema.europa.eu/partners-networks/networks/european-network-paediatric-research-european-medicines-agency-enpr-ema

[43] Assent is a term used to express willingness to participate in research by persons who are by definition too young to give informed consent but who are old enough to understand the proposed research in general, its expected risks and possible benefits, and the activities expected of them as subjects.

[44] Informed Consent for Paediatric Clinical Trials in Europe 2015, EnprEMA (European Network of Paediatric Research at the European Medicines Agency); https://www.ema.europa.eu/documents/other/informed-consent-paediatric-clinical-trials-europe-2015_en.pdf

[45] eYPAGnet (European Young Person's Advisory network) is part of the umbrella organisation, the International Children's Advisory Network (iCAN) and has the following goals: (i) to improve the capacity of collaboration with the different actors, who participate in the research and development process of innovative drugs; (ii) to gather a variety of experience related with different pathologies; (iii) to promote the planning and development of clinical research initiatives for children, on a European level; (iv) to consolidate the curriculum of capacity-building and empowerment training programs to the young patients; (v) to promote and lead the creation of new chapters; and (vi) to empower the selection of professional careers in the scope of science, among the youth.

4.4.4. Common themes across disease areas

Some common themes were articulated by all patient groups. Firstly the nature and complexity of consent forms remain a significant concern. Trust is eroded if, for example, the consent fails to mention use of the data beyond the original research objectives and nevertheless data is still ultimately shared. By contrast, trust is engendered and consent often given willingly by properly explaining the plans for sharing data with third parties. Of course, assurance of comprehension of the consent itself remains a challenge.

Hence all patients whatever their vulnerability, age or illness should have clear, readable, and widely available information on data protection across multiple media sources; be able to give informed consent before their personal data are collected; be briefed about the purpose for collecting data; have protection over sensitive data; be able to control the level of personal data collected and shared; have the reassurance that patients' experience is represented in decision-making about data; and be assured about prevention of abuse or misuse of personal data. Importantly, physicians and other professionals interacting with patients should have basic knowledge about data use and be able to explain to patients the limits and benefits of data sharing.

Patient representatives universally agreed that consent forms should be simplified and the length reduced. Terminology should be accessible and understandable, and key information regarding opt out possibilities and data sharing, particularly around its use for secondary research purposes, should be presented in a separate section from the legal information. While there was recognition that the residual risk of re-identification can never be zero especially in an era of increasing cyber-attacks on data (e.g. WannaCry ransomware cyber-attacks which crippled among others the UK National Health Service networks) which may compromise data privacy, the risk needs to be transparent and clearly communicated alongside the safeguards in place to protect their data such that patients understand the potential consequences of data sharing. Building trust was key in these interactions. Lastly the ethics of differential patient opinions were discussed; how could processes other than consent allow for individuality of decision making while ensuring the representativeness of the dataset? Lastly, challenges remain as to how to share historical data collected under broad consent, no longer supported under GDPR, but where data linkage to other datasets may deliver novel insights.

5. The Mechanics of Anonymisation – meeting the challenge of different datatypes

- *To define strengths and limitations of current methodology while both maximising scientific utility and considering the international perspective.*
- *To illustrate challenges of the technological approaches with concrete case examples.*
- *To discuss whether anonymisation techniques are equivalent across different datasets.*

5.1. A review of anonymisation techniques – strengths and limitations of different methods across different jurisdictions

Legislation in the US and in the EU protects individuals from being identified from their health information; in both jurisdictions, once the data have been anonymised, processing of the data no longer falls under the scope of the legislation. US and EU regulations differ, however, in a fundamental manner; HIPAA describes two potential methodologies to achieve de-identification while GDPR does not specify particular approaches to render data anonymous.

HIPAA offers two strategies for anonymisation, safe harbor and expert determination. The former requires the removal of 18 quasi-identifiers but the data controller must certify that he or she has no 'actual' knowledge that the residual information could be used to identify an individual; however, a controller’s 'actual knowledge' is extremely hard to police. Expert determination requires the application of statistical or other scientific methods and the subsequent demonstration that a very small risk is anticipated that a recipient could re-identify an individual. However, HIPAA does not clearly define 'appropriate knowledge and experience' nor does it specify what an acceptable level of risk of identification is. The EU legislation is deliberately non-prescriptive regarding methodology but is clear that in determining whether "a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly" (Recital 26 of EU GDPR).

A celebrated US case (Sweeney, 1997) illustrates how anonymisation can be overcome by combining two datasets. Following Governor William Weld’s hospitalisation after collapsing on stage, Latanya Sweeney used hospital discharge data together with voter list data coupled with the approximate time of hospital admission to identify and access the governor’s health records. The elements common to the two datasets were the ZIP (postal) code, birthdate and gender. While recognising that this example occurred now nearly 20 years ago and before HIPAA came into effect, it illustrates the principle that to preserve privacy, the first step must be to carefully characterise the dataset. In any dataset there will be clear identifiers which must be removed, redacted, transformed, or replaced. However, the handling of quasi-identifiers requires careful balancing of scientific utility with the risk of identification; utility may be lost if all quasi-identifiers are removed.

So how would the balance between personal privacy and data utility be constructed? A pragmatic approach to anonymisation was proposed that involves three steps: first the creation of a 'gold standard' dataset, for example an appropriate selection of records, in which identifiers and quasi-identifiers are annotated by a number of different investigators to propose and create rules; second, using a machine to apply the rules to the entire dataset; and finally, measuring the effectiveness of automatic detection of identifiers. The measures for effectiveness of anonymisation are: the rate at which real identifiers were detected ('recall', R); the rate at which the detected identifiers were in fact real identifiers ('precision', P); and the weighted average of recall and precision (F). This informs on risk of re-identification but does not provide an absolute level of risk as it cannot account for re-identification through linkages to other datasets.

As was emphasised repeatedly throughout the workshop, no approach is absolute. One should assume that some potential identifiers will be leaked; hence it is important to identify the potential for re-identification based upon the information that was leaked. It may be minimal despite the information leak and hence sufficiency of anonymisation would still be achieved.

In order to make the system more robust, a preferable approach may simply be to replace names and explicit identifiers with pseudonyms rather than redact them. In this way any personal identifiers, which may have been missed by automated strategies, would be less obvious. Identifiers would effectively be 'Hiding in Plain Sight' (Carrell et al, 2013). But with this relatively new technique, the possibility that a computer is able to mimic the initial 'detect-and-replace' strategy, redact the fake identifiers, and increase the vulnerability to identification remains possible (Li et al, 2017). The means to reduce this vulnerability comes with a massive loss in the precision of the technique, i.e. replacement of significant amounts information that are not identifiers (Li et al, 2017) which itself markedly reduces utility. Moreover, this technique cannot address the semantics of the text and hence may miss words which could impart meaning but would not be identified by an automated approach.

Yet another method to anonymise data is termed 'k-anonymity'; in this model, 'k' represents the number of individuals in a dataset who share a set of characteristics (such as age, gender, postal code) that could be used to identify them. Generalising the characteristics (e.g. using age bands or making postal codes less specific), or even removing the characteristics altogether, will mean that several individuals (numbering k) cannot be distinguished from each other, thus reducing the risk of re-identification. However, the possibility of re-identification increases if an adversary attacks several records in the dataset.

All these techniques for anonymisation – including redaction, hiding in plain sight, and k-anonymisation, are vulnerable to attack and involve a trade-off between privacy and utility. Redaction, while feasible and potentially easy, leaves residual identifiers readily discoverable unless the redaction is so extreme as to reduce utility markedly. Hiding in plain sight may help but needs to be applied in a manner that mitigates attacks against it, that itself may reduce the precision of anonymisation and ultimately data utility. Finally, while one can model privacy and utility separately it may be better to manage them simultaneously. Importantly researchers using different datasets (e.g. for linkage or meta-analysis) need to know what anonymisation methods have been used in order to combine the data and reach meaningful conclusions.

5.2. Comparison of anonymisation techniques in the context of clinical study reports

When pharmaceutical companies send anonymised clinical study reports (CSRs) to EMA in the context of Phase 1 of Policy 0070, an anonymisation report is also required, which must include information on and justification of the anonymisation method used and an analysis of the risk of re-identification. As of October 6th 2017 and as presented at the workshop, 54 anonymisation reports have been published, 11 of which addressed CSRs without patient identifiers and 43 which addressed CSRs which contained identifiers, 9 of which were in the orphan field. In 8 of these applications, redaction was the anonymisation technique applied and in 7 of those cases there was full redaction of the case narratives. In only 2 cases was the redaction applied selectively to elements such as demographic characteristics, medical history, and verbatim text. A similarly conservative approach was employed for adverse events (AEs): in 3 cases AEs were redacted completely, and in 3 cases only adverse reaction case narratives and in-text narratives were redacted. Hence where the risk of re-identification is considered highest (small samples sizes and innovative products, e.g. gene therapy) a conservative redaction approach is chosen which in the majority of cases has led to a full redaction of the case narrative.

Clinical trials for medicines for rare diseases are often are conducted, by necessity, with small populations and inclusion criteria may be on the basis of rare genetic mutations or other biomarkers which increases vulnerability to data attacks. Even with non-orphan products containing larger studies (>100 participants), 97% of studies employed redaction as the anonymisation technique. In 90% of cases (26/29 applications), a qualitative risk assessment was performed based only on a subjective evaluation and in none of these studies was the risk of re-identification determined. There was some heterogeneity in the extent of redaction applied but in 50% of cases, full redaction of the case narrative occurred; only 15% had selected redaction of some study categories. For orphan applications, extensive redaction of AEs occurred: 31% redacted all AEs with only 9 applications applying selectivity. For the few applications employing a quantitative assessment, a risk threshold was set (0.09), the risk of re-identification was calculated, and the assumptions made were less conservative than in the qualitative approach.

Analysis of the few anonymisation reports received to date suggest that, on the whole, the impact of a full redaction of the narrative was poorly addressed. In the limited experience of EMA to date, the disease and/or study population appear to be correlated with, and perhaps driving, the choice of anonymisation process, with no clear specification as to why. Confidence was limited within companies in the assumptions being made for assessing risks.

5.3. Anonymisation techniques in the context of individual patient level data

The experience and practices of data sharing by two large pharmaceutical companies were presented. The experience of sharing clinical trial data voluntarily through the ClinicalStudyDataRequest.com (CSDR)[46] was compared with that of Policy 0070; these platforms differ in how data is shared, the safeguards in place, and, consequently, anonymisation strategies employed.

CSDR allows access of clinical study data for legitimate research proposals through a terms-of-use agreement and via a secure portal; the secondary researcher commits to publish an analysis of the data. External measures thus enable a minimum level of anonymisation that nevertheless meets data protection obligations and provides a high level of data utility. Furthermore by defining an anonymisation strategy which meets global standards, it can be utilised across studies ensuring a consistent approach but also the ability to retain data relationships. Anonymisation is iterative with cycles of anonymisation and risk-testing until the risk of identification falls below a chosen threshold. Such approaches, however, are more burdensome and have higher resource implications.

A different approach is taken by sponsors for compliance with Policy 0070. Defined sections of the CSRs and clinical summary documents of applications submitted to the Committee for Medicinal Products for Human Use (CHMP) for authorisation, even if subsequently withdrawn, are shared via the Phase I of Policy 0070. The company complies with anonymisation obligations by redacting patient narratives and per-patient per-visit line listings but several companies are now exploring a risk-based approach, as an alternative to redaction, to better balance the risk of re-identification against data utility. Assumptions used for estimating the risk of identification include the motives and knowledge of adversaries, the threshold for an acceptable level of risk, definition of a similar population, technical limitations of techniques (e.g. hiding in plain sight), and the data context in relation to quasi-identifiers. Assumptions around attackers and their motives are typically qualitative and it is envisaged that it is possible to incorporate such assumptions into a quantitative assessment of risk.

Knowledge about the risk of re-identification and the interpretation of risk assessments is still evolving. It is important to be cognisant of the fact that the reports posted in compliance with Policy 0070 at least will be in the public domain for many years and thus the risk of re-identification will change as methods of re-identification and data science evolve.

While contributors emphasised the benefits that could be gained for all stakeholders by meaningful data sharing, existing regulations and guidelines leave multiple uncertainties in sharing data: clear, worldwide, recognised guidelines are urgently needed. Data anonymisation is necessary for not only for multi-stakeholder, external data sharing but also internally within a multinational company. Creation of a single data warehouse that hosts all clinical research data (both internally generated

[46] ClinicalStudyDataRequest is a consortium of clinical study sponsors/funders committed to share high quality patient level data from clinical studies through the CSDR platform. CSDR facilitates this aim by creating a research friendly platform utilising industry leading practices, including an independent review of proposals and protection of patient privacy and confidentiality; https://www.clinicalstudydatarequest.com/Default.aspx

(clinical trials, pharmacovigilance) and externally acquired (patient cohorts and real world studies) allows access to all company employees. Although the level of anonymisation needed might be considered different for internal and external use, it was proposed that the same most stringent criteria should be applied to both contexts of release, with the principal difference being the need for review and approval by either an independent expert or review panel when sharing data externally. In order to facilitate the timeliness of data sharing, anonymisation of the entire study database at the time of the completion of the clinical study report appears to be most efficient.

5.4. Does one size fit all? – Challenges of anonymising real world data

In the big data era, a wider variety of data will increasingly be captured in the context of clinical trials. Such data brings different challenges to those from clinical study data not only in terms of the structure and standardisation of the data but also in its content. As such, the scope of the workshop was extended to include real world data from patient registries and individual cohort studies. The experience of Public Health England (PHE, a government body to protect and improve citizens' health and wellbeing and reduce health inequalities) illustrated the challenges around the use and processing of health data generated through the process of delivering normal clinical care, i.e. real world data.

In the UK, a unique legislation (Section 251 of the UK NHS Act 2006), provides specific legal permission to collect information about patients without the need to seek consent where the information is needed to support essential NHS activity. This legislation therefore enables the collection of data, within a national population registry, from about 400,000 newly diagnosed cancer patients and patients for whom a cancer diagnosis is suspected. As the depth and detail of the individual patient journey captured by the registry is rich, that then mandates strict policies to govern data collection, storage and release. However, the data are of considerable value for a range of activities including service planning, clinical audit by individual clinicians, and the development of prognostic disease models. PHE is therefore exploring novel mechanisms by which data can be shared while respecting data privacy. Approaches include data aggregation (e.g. releasing annual counts of cancer diagnosis and suppressing small cell values), low-level release (e.g. involving restricted access to specific data), and releasing synthetic data identical to real data (The Simulacrum) [47].

The Simulacrum will model as many of the properties of the original data as possible but contains no real patient data. As such, ultimately the Simulacrum will enable researchers to generate and test hypotheses, refine research questions, and test feasibility; final queries generated through the artificial data can then be validated on the original data in a secure environment. The latest iteration of Simulacrum contains complex linked data, patients with multiple tumours, and detailed chemotherapy treatment data.

Many real world datasets offer the opportunity to link to additional datasets via unique patient identifiers that can bring depth and richness to the analysis, but the visibility of the linkages in addition to potential patient specific identifiers, potentially increases the risk of re-identification. 'OpenPseudonymiser' is an open-source software created by the University of Nottingham[48] that allows safer data linkages. The software replaces each patient identifier at source with a pseudonym derived from their unique health service number allowing researchers to reliably link datasets (that have been pseudonymised in identical ways) but still not have access to patients' demographic information. Use

[47] https://healthdatainsight.org.uk/project/the-simulacrum/
[48] https://www.openpseudonymiser.org/

of the software can be extended to cover data beyond patient identifiers using per patient encryption keys but sharing depends on trust between researchers to prevent aggressive re-identification.

One size does not fit all: anonymisation decisions are task specific and choices will always be needed around which axes of data fidelity to preserve. Domain knowledge is essential in allowing a privacy impact assessment. Data release will always be a risk-balancing exercise, especially rich real world data that contains detailed patient encounter history. Simulated data approaches such as being developed by Simulacrum, however, provide a pragmatic solution that allows complex queries to be developed and hypotheses refined and yet protect patient privacy. Lastly for real world data, where value may lie in linkage with other datasets, pseudonymisation approaches allows linkages; retention of an identifying code nevertheless increases the risk of re-identification and therefore data sharing models rely on trust between researchers and external controls (e.g. data use agreements, secure compute sites). Thus, combining approaches such as OpenPseudonymiser with data sharing agreements may mitigate the risk. Future work is required to develop reproducible algorithms for linking anonymised and simulated data to other real and simulated datasets.

6. Balancing access and data utility

- *To define how different mechanisms of access (from open access to a range of controlled access solutions) influence anonymisation approaches and ultimately data quality.*
- *To consider challenges for operationalising clinical data sharing.*
- *To discuss challenges for accessing and analysing data from the user perspective.*

6.1. Overview of data sharing possibilities to facilitate international data sharing

The objective of data sharing platforms is to serve as rich sources of valuable clinical data to facilitate new science and health research. Data-sharing platforms must balance conflicting needs: maximising access to researchers while preserving privacy; maximising utility while employing anonymisation techniques. Data-sharing platforms themselves should deliver an environment that simplifies the administrative process and provides analytic possibilities to enable new science and discovery.

Measures for safeguarding the data vary across platforms, ranging from the sole use of anonymisation methods (e.g. Project Data Sphere[49]) similar to those on the EMA data-sharing platform to a controlled data access where data are supplied only in a closed, secure system (e.g. CSDR, Yoda Project[50], Soar[51]) or where data are available either to download or within a closed and secure system depending on the data contributor and other factors (e.g. Vivli[52]). However, it is not clear if these measures provide meaningful barriers against malicious or inadvertent breaches of confidentiality. What needs to be clarified is whether the more rigorous and burdensome measures safeguard and deter breaches of confidentiality or only create unnecessary barriers for scientific progress and collaboration.

The following points are pertinent to data sharing, particularly in a global context:

[49] https://projectdatasphere.org/projectdatasphere/html/home
[50] http://yoda.yale.edu/
[51] https://dcri.org/about/who-we-are/our-approach/data-sharing/soar-data/
[52] https://vivli.org/

- many trials are multi-regional; data providers must ensure that data provision complies with country-specific and local legislation and policies;
- anonymisation methods must comply with global (and local) principles;
- data-sharing platform's technology and legal infrastructure should have global reach;
- the platform's ability to link multiple types of datasets could increase the risk of re-identification;
- in the future platforms should be ready to support the sharing and integration of more 'challenging' data such as imaging, genomics and real-world data.

Each stakeholder brings a unique perspective. From the platform perspective, responsibility for data anonymisation rests with the data contributor, not the platform. The ability to link that anonymised dataset, however, with other datasets through the platform increases the risk of re-identification,. Given that the data contributor will have anonymised the data in the context of safeguards provided by the platform, some accountability for data protection must lie either with the platform or the subsequent data requestor. Special provision is needed for sensitive datasets (e.g. those on rare diseases) where balancing anonymisation and data utility is particularly challenging.

Many data contributors undertake anonymisation only at the time of data request. This 'just in time' approach often results in significant delays in providing access to the data. Industry data providers weigh the expense of the anonymising data in-house or outsourcing the task; for academic data providers the financial and human resources required present particular barriers for academic-led trials. Allocation of specific funding for anonymisation as part of the trial funding in combination with access to appropriate expertise would help to promote a data sharing culture. Additionally, funders of clinical trials, whether academic, for-profit, foundation, or other non-profit organisations, should require data sharing as a requirement for funding and include the resources required within the funds allocated.

Initial evidence suggested that requests for access to data already available and resulting publications have been disappointing low (Strom et al, 2016), potentially secondary to requirements to analyse data behind a firewall, challenging meta-analyses of patient-level data that comes from different sources. However, more recent data suggests demand is increasing in addition to the number of publications arising from the data shared (Coady et al, 2017; Ross et al, 2018). Data interoperability, impacting on the ability to combine datasets, is a further challenge, stemming from the absence of global data and anonymisation standards. Such standards should address anonymisation methodologies, quality standards, approaches to risk assessment, and a standard for transparency.

Ultimately cultural change is needed to drive behaviour change supported by data sharing processes that preserve data utility and personal autonomy at minimal burden to data contributors. Importantly, individuals who elect to share the data they have generated should be recognised for their academic contribution, in much the same way as for peer reviewed publications. Appropriate metrics for data sharing activities, accepted by funding bodies and academic institutions, need to be developed to assign recognition. Undoubtedly academic recognition for sharing datasets will do much to encourage and facilitate a data sharing culture.

7. Future challenges for data anonymisation

- *To consider how anonymisation approaches can keep pace with the evolving scientific landscape.*
- *To consider what additional challenges will be posed by linking multiple datasets, including genomic and healthcare data, and the challenges raised by new innovative datasets.*
- *To discuss how anonymisation approaches can be future-proofed.*

7.1. Influence of changing scientific landscape on data protection

Health data exists in a fast-evolving scientific ecosystem. The standard data sources comprise data from health services (e.g. clinical records, prescribing and laboratory tests), public health (e.g. population statistics and disease surveillance) and research (e.g. clinical trials, registries, and biobanks) but increasingly these data sources will interact with data from environmental, sociological, behavioural, lifestyle, and socioeconomic domains. Critical to the exploitation of these data are increasing technological capabilities that allow the use of the data in new ways (technological, analytical, and policy) and a wide array of stakeholders (individual citizens and groups that represent them, health services, research and academic communities, healthcare industry, data and information and communication technology industry, and government). This is a dynamic, organic ecosystem with strong interconnections and interdependencies amongst the various actors, processes, and data sources in which changes in one part of the ecosystem impacts on another.

There is a high level of interdependence between the different segments of this scientific ecosystem. Advances in capabilities allow use of data in new ways and lead to new discoveries which in turn can be used with demographic data to improve and hasten diagnosis. For this type of processing, safeguards can allow storing patient data securely with anonymisation and encryption. The challenge lies in balancing the use of innovative capabilities on one hand with patient privacy on the other.

While innovation and technology bring major challenges, it can also bring insights and solutions. The proliferation of patient platforms and distributed data systems hosting personal data brings challenges around data mobility and the distributed responsibilities in such a framework. This, when coupled with the vision for data portability as laid down in Article 20 of the GDPR, will have implications for data access and also for data protection. Global guiding principles are urgently needed to address the tension created between supporting innovation in a dynamic and changing data environment and the need to protect personal data. Such principles need to incorporate a number of key features (Stilgoe et al, 2013; Vayena & Blasimme, 2018):

Anticipatory

- **Adaptivity**

  Readiness to devise specific oversight mechanisms for new data types and uses

- **Flexibility**

  Treating data according to their actual use rather than on their source

- **Monitoring**

  Continuous monitoring process to detect signals of new vulnerabilities for data subjects

Monitoring of activities of stakeholders

Responsiveness

Preventing vulnerabilities from resulting in actual harms

Ensuring effective containment of harmful effects in case of failures: e.g. privacy breaches

Promoting responsive innovation

Reflexivity

Consider the effects of new practices as they emerge from data-driven research

Critically appraise assumptions embedded in research practices and regulatory mechanisms

Cultivate awareness for governance models affect rights and interests of individuals and communities

Inclusiveness

Upstream engagement of relevant stakeholders, including lay public, before technological path dependences become established

Transparency

Openess and transparency of governamce, process, analytics to foster trusted relationships

Such principles should not create barriers but rather enable data sharing by creating an environment of trust among relevant stakeholders. Anonymisation is one critically important approach but should be coupled with other safeguards to maximise that balance of data utility with personal privacy.

7.2. Technological solutions for data anonymisation

The concept of artificial intelligence (AI) has caused a sea change in thinking and has already delivered remarkable results, as illustrated by the success of AlphaGo challenge in 2016, a game previously considered to be too complex and difficult for AI. The hope is that, as these approaches mature, they will derive insights from complex and heterogeneous data sources not achievable by traditional analytics. However, current law regulating personal privacy is based on an understanding of how humans process information, especially how humans remember and forget (Villaronga et al, 2017). Thus, novel deep learning approaches across multiple datasets not only threaten anonymisation but also potentially make it impossible to fulfil the legal objectives of the 'right to be forgotten' (Villaronga et al, 2017). Techniques that protect personal information yet enable innovative analytical approaches are needed.

In fact, AI itself may offer a new approach to anonymisation by generating artificial data on which algorithms and models can be tested; a similar approach to that of the Simulacrum. Projects such as the Synthetic Data Vault (Patki et al, 2016) have demonstrated that synthetic data can successfully replace original data for data science investigations. Such approaches are likely in the future with the development of quantum computers where the additional storage and computing power allows solutions to problems to be explored simultaneously rather than sequentially. Synthetics databases will not be suitable across all scenarios, however, and a number of other methodologies offer possible analytical solutions that can operate across multiple data assets; quantum cryptography, secure multiparty computation, polymorphic encryption, and pseudonymisation, each with their own inherent advantages and disadvantages.

Secure multiparty computation allows a set of parties not bound by trust agreements to perform the computation in a distributed manner, while each remains oblivious to the input data and the intermediate results. As such, the computation is secure and ultimately each party is aware only of its own input and the results. The SODA (Scalable Oblivious Data Analytics) project[53] is based on this approach and therefore does not ask data subjects and controllers to share personal information, but only to make it available for encrypted processing, often within the firewall of the data controller. The challenge will be ensuring such approaches still meet the needs of the end user.

Homomorphic encryption allows computations on encrypted data, generating an encrypted result which when decrypted, matches the results of the operation performed on the non-encrypted data and as such may allow secure but distributed processing of clinical data. However, encryption processes are slower than computing on the original data, may introduce noise, and may limit the number of operations that can be performed. New developments such as partial homomorphic encryption may address some of these limitations and is the approach adopted by the MyHealthMyData project[54].

These are promising approaches but multiparty computation and homomorphic encryption are reversible and are thus currently considered pseudonymisation techniques. Opinion 05/2014 on Anonymisation Techniques, issued by the Article 29 Working Party[55] points out that anonymisation must be irreversible. Moreover, according to the Article 29 WP, "a specific pitfall is to consider pseudonymised data to be equivalent to anonymised data (…); importantly pseudonymised data cannot be equated to anonymised information as they continue to allow an individual data subject to be singled out and linkable across different data sets". With this in mind a number of anonymisation techniques including noise addition, differential privacy, permutation, aggregation/K-anonymity and L-diversity/T-closeness need to be reassessed.[56] However, were the encryption key cancelled, or the initial identifiable data destroyed, an adequate anonymisation standard may be reached.

A remaining question is whether 'qualified anonymity' may be acceptable; here data undergo anonymisation techniques (such as homomorphic encryption and secure multiparty computation) which renders them (i) pseudonymised for the hospital, the sole entity holding the re-identification key (e.g. for fulfilling the duty of care) and (ii) anonymised for any third party receiving the dataset. Such an approach would meet researchers' needs to retain the capacity for re-tracing and singling-out specific participants in a study in order to, for example, assess the progression of disease and the long-term outcomes of treatments, or simply to keep them informed about unexpected findings or life-saving discoveries. It was argued that the ability to identify individuals is not only something that may happen, rather it is something that must happen, but only under specific circumstances defined as a proper 'qualification' by the law (e.g. judges fulfilling their official duties, researchers finding a cure that may eradicate a disease, unexpected findings which are potentially curable, public authorities exercising their powers, etc.).

Innovation and future research development needs a market capable of extracting the value of healthcare data. The paradox that must be solved, however, is that in some cases in the EU commercial transactions on data are lawful only if the data are anonymised; the 'specific' consent and re-consent requirements required by pseudonymisation may be impractical and possibly highly counterproductive. Solutions are urgently needed: MyHealthMyData aims to guarantee privacy and

[53] https://www.soda-project.eu/
[54] http://www.myhealthmydata.eu/
[55] The technical body tasked with providing the European Commission with independent advice on data protection matters and supporting the development of harmonised policies for data protection.
[56] Opinion 05/2014 on Anonymisation Techniques, par. 5.2.

security of healthcare data by introducing a distributed architecture based on blockchain and smart contracts. The project is in its early stages but aims to develop a comprehensive methodology to guide the implementation of data and identity protection systems, specifically defining approaches and tools to classify sensitive data based on medical as well as predictive, and potentially economic, value. The project will also analyse users' behavioural patterns alongside ethical and cultural orientations.

8. Conclusions

The constantly changing scientific, technological, and legislative landscape is challenging our ability to share data in sufficient depth and detail to maintain its scientific utility while meeting data protection obligations. Additional challenges are posed by what is now a global data environment spanning multiple legislative jurisdictions. Global guiding principles, including international agreement on common terms, are urgently needed to address this tension and find appropriate balance to derive the benefits of data sharing. Further, a commitment to transparency on the choice and methodology of anonymisation approaches is necessary to facilitate international clinical data sharing.

There was agreement that no data should, a priori, be exempt from data sharing but that a risk assessment should be performed based on a framework for anonymisation to determine the risk of re-identification and if, given the specific context, that risk were acceptable. The risk assessment should incorporate the sensitivity of the data and the context of release coupled with methodologies to validate the risk assessment calculations. The need to maintain the scientific utility of the data may also influence the choice of anonymisation approach. Furthermore new legislative requirements for the periodic evaluation of anonymisation in the light of a continually changing data environment necessitate associated metrics to allow monitoring of the process and bring accountability for all involved stakeholders. Where data cannot be shared openly, strong data sharing agreements and other external measures are essential. Lastly, future innovation and research development will extract value from data in an evolving scientific landscape, thus requiring continuous and proactive exploration of novel anonymisation approaches.

Importantly, where consent is used as a legal basis for data sharing, there must be clarity as to what will happen to the data, how it will be used, and how the rights of the data contributor will be met both at data disclosure and over time. Engagement with all stakeholders is necessary to communicate the benefits of data sharing but also to reassure that the risk of re-identification is acceptably low. Ultimately cultural and behavioral change, coupled with mechanisms to enable safe data sharing, reduce burdens on the data contributors, and provide a process that is not overly onerous for all stakeholders are necessary. Importantly, we emphasise that those who share the data they have generated should receive academic recognition for the contribution, paralleling the contributions of peer reviewed publications; data sharing metrics, accepted by funding bodies and academic institutions, need to be developed to assess the value of the data shared and the impact of the data sharing activities.

The principles discussed throughout the workshop will form the basis of later communications in order to reflect developments which have occurred since the meeting. It is important that such principles reflect current best practices in anonymisation and recognise recent significant activity.

The sensitive and personal nature of healthcare data demands robust data protection combined with careful communication of the potential benefits of data sharing for the public good. The benefits of data sharing should not come at the cost of privacy. A clear framework of anonymisation able to meet the varied global legislative requirements that quantifies the risk of re-identification both at the time of data release and into the future is essential to build patient trust and confidence in the accountability of the process.

References

Bierer, B.E., Crosas, M., Pierce, H.H. (2017) Data authorship as an incentive to data sharing. New Engl J Med; 377(4): 402.

Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark,C., Wellner, B., Hirschman, L. (2013) Hiding in plain sight: use of realistic surrogates tio reduce exposure of protected health information in clinical text. J Am Med Inform Assoca; 20: 342–348.

Coady, S.A., Mensah, G.A., Wagner, E.L., Goldfarb, M.E., Hitchcock, D.M., Giffen, C.A. (2017) Use of the National Heart, Lung and Blood Institute Data Repository. N Engl J Med; 376: 1849–1858.

Danezis, G., Domingo-Ferrer J., Hansen, M., Hoepman, J-H., Le Métayer, D., Tirtea R., Schiffner, S. (2014) Privacy and Data Protection by Design – from policy to engineering. www.enisa.europa.eu

Danker, F.D., El Emam, K., Neisa, A., Roffet, T., (2012) Estimating the re-identification risk of clinical data sets. BMC Medical Informatics and Decision Making, 12:66 (doi:10.1186/1472-6947-12-66).

El Emam, K., (2012) Guide to the De-Identification of Personal Health Information. CRC Press.

El Emam, K, Jonker, E., Arbuckle, L., Malin, B. (2011) A systematic review of re-identification attacks on health data. PLoS One; 6:e2807:10.1371/journal.pone.0028071.

Elliot, M and Mackey, E., (2014) The Social Data Environment. Digital Enlightment Yearbook, 253–263.

Gymrek, M., McGuire, A.L., Golan,D., Halperin,E., Erlich,Y. (2013). Identifying personal genomes by surname inference. Science; 339: 321–324.

Kosinski, M., Stillwell, D., Graepel, T. (2013) Private traits and attributes are predictable from digital records of human behaviour. PNAS; 110:5082-5805.

Lasalvia, A., Zoppei, S., Van Bortel, T., Bonetto, C., Cristofalo, D., Wahlbeck, K., Bacle, S.V., Van Audenhove, C., van Weeghel, J., Reneses, B., Germanavicius, A., Economou, M., Lanfredi, M., Ando, S., Sartorius, N., Lopez-Ibor, J.J., Thornicroft, G.; ASPEN/INDIGO Study Group. (2013) Global pattern of experienced and anticipated discrimination reported by people with major depressive disorder: a cross-sectional survey. Lancet; 381(9860): 55–62.

Lepola, P., Needham, A., Mendum, J., Sallabank, P., Neubauer, D., de Wildt, S. (2016) Informed consent for paediatric clinical in Europe. Arch Dis Child 2016; 101: 1017–1025.

Li, B., Vorobeychik, Y., Li, M., Malin (2017) Scalable iterative classification for sanitizing large scale datasets. IEEE Trans Knowl Data Eng; 29: 698–711.

McCormack, P., Kole, A., Gainotti, S., Maxcalzoni, D., Molster, C., Lochmuller, H., Woods, S. (2016) 'You should at least ask'. The expectations, hopes and fears of rare disease patients on large-scale data and biomaterial sharing for genomics research. Eur. J Hum. Genet; 24: 1403–1408.

Mello, M.M., Francer, J.K, Wilenzick, M, Teden, P, Bierer, B.E, Barnes, M. (2013) Preparing for responsible sharing of clinical trial data. New Engl J Med; 369: 1651–8.

Mello, M.M., Van Lieou, B.S., Goodman, S.N. (2018) Clinial trial participants' views of the risks and benefits of data sharing. New Engl J Med; 378: 2202–11.

Patki, N., Wedge, R., Veeramachaneni, K. (2016) The Synthetic Data Vault. IEEE International Conference on Data Science and Advanced Analytics (DSAA); 399–410.

Robertson, J. States' hospital data for sale puts privacy in jeopardy. Bloomberg News (2013). Jun 5. www.bloomberg.com/news/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy.html.

Ross, J,S., Waldstreicher, J., Bamford, S., Berlin, J.A., Childers, K., Desai, N.R., Gamble, G., Gross, C.P., Kuntz, R., Lehman, R.L., Lins, P., Morris, S.A., Ritchie, J.D., Krumholz, H.M. (2018) Overview and experience of the the YODA Project with clinical trial data sharing after 5 years. Scientific Data; 5: 180268.

Stilgoe, J., Owen, E., Macnaghten, P. (2013) Developing a framework for responsible innovation. Res Policy; 42: 1568–1580.

Strom, B.L., L. Buyse, M.E., Hughes, J. Knoppers, B. (2016). Data Sharing — Is the Juice Worth the Squeeze?. New Engl J Med; 375: 1608-1609.

Sweeney, L. (1997) Weaving technology and policy together to maintain confidentiality. Joumal of Law, Medicine & Ethics, 25 (1997): 98–110.

Sweeney, L. Matching known patients to health records in Washington State data. Harvard University, Data Privacy Lab (2013).

Taichman, D.B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., Hong, S-T., Haileamlak, A., Gollogly, L., Godlee, F., Frizelle, F.A., Florenzano, F., Drazen, J.M., Bauchner, H., Baethge, C., Backus, J. (2017) Data sharing statements of Clinical trials: A requirement of the International Committee of Medical Journal Editors. Annals Int Med; 167: 63–65.

Thornicroft, G., Brohan, E., Rose, D., Sartorius, N., Leese, M. INDIGO Study Group. (2009) Global pattern of experienced and anticipated discrimination against people with schizophrenia: a cross-sectional survey. Lancet; 373(9661): 408–15.

Vayena, E., Blasimme, A. (2018) Health Research with Big Data: Time for Systemic Oversight. J Law Med Ethics; 46: 119–129.

Villaronga, E.F., Kieseberg, P., Li, T. (2017) Humans forget, machines remember: Artificial Intelligence and the right to be forgotten. Computer Law & Security Report. DOI: 10.1016/j.clsr.2017.08.007