The conflict between privacy and scientific research in the GDPR

Janos Meszaros
Academia Sinica
postdoctoral researcher
Taipei, Taiwan
dr.janos.meszaros@gmail.com

***Abstract*— One of the most important goals of the General Data Protection Regulation (GDPR) is to protect the data subject's privacy in the online environment where there is a significant imbalance between the users and the data controllers. To achieve this goal, the GDPR requires stronger consent to protect the data subjects and introduced new rights, such as the right to be forgotten and data portability. The new and stronger rights in the earlier drafts of the GDPR would have significantly restricted the data processing activities not just for the online services (e.g., Google, Facebook, Twitter), but for all the other data controllers. Furthermore, the planned strict rules could have been a considerable burden, especially for the entities processing sensitive data, such as research institutes and pharmaceutical companies.**

It became clear at the later stages of the drafting period that the new rules aiming at the online environment could also hamper the scientific research and technological developments. The analysis of large datasets (big data) seemed also challenging by the earlier drafts of the GDPR.

A compromise was necessary between the data subject's new and stronger rights and the researchers' interests if the EU did not want to fall back with the innovation, especially in the field of technological developments and biomedical research. These reasons led to the involvement of pseudonymisation, statistical and scientific research exemptions in the final version of the GDPR.

Keywords— privacy, scientific, research, big data, GDPR

I. INTRODUCTION

The GDPR is a significant step in the EU to shift from the era of "small data" to big data with several exemptions for the scientific research and statistical purposes. The GDPR opens up the possibilities for using big data in research, but the burden remains on the EU member states to elaborate the details to set the opportunity in action. To interpret the normative text of the GDPR and to highlight its acceptance toward big data and scientific research, it is crucial to examine the earlier proposals and the relationship between the Recital and the normative text of the Regulation.

Viktor Mayer-Schönberger and Yann Padova examined three privacy aspects of big data: data collection, repurposing and longer retention periods [1]. These are the key areas of data protection law to allow the use of big data in research. However, allowing research and technological developments in these fields may also have a negative effect on the data subjects' privacy.

II. BIG DATA AND SCIENTIFIC RESEARCH IN THE GDPR

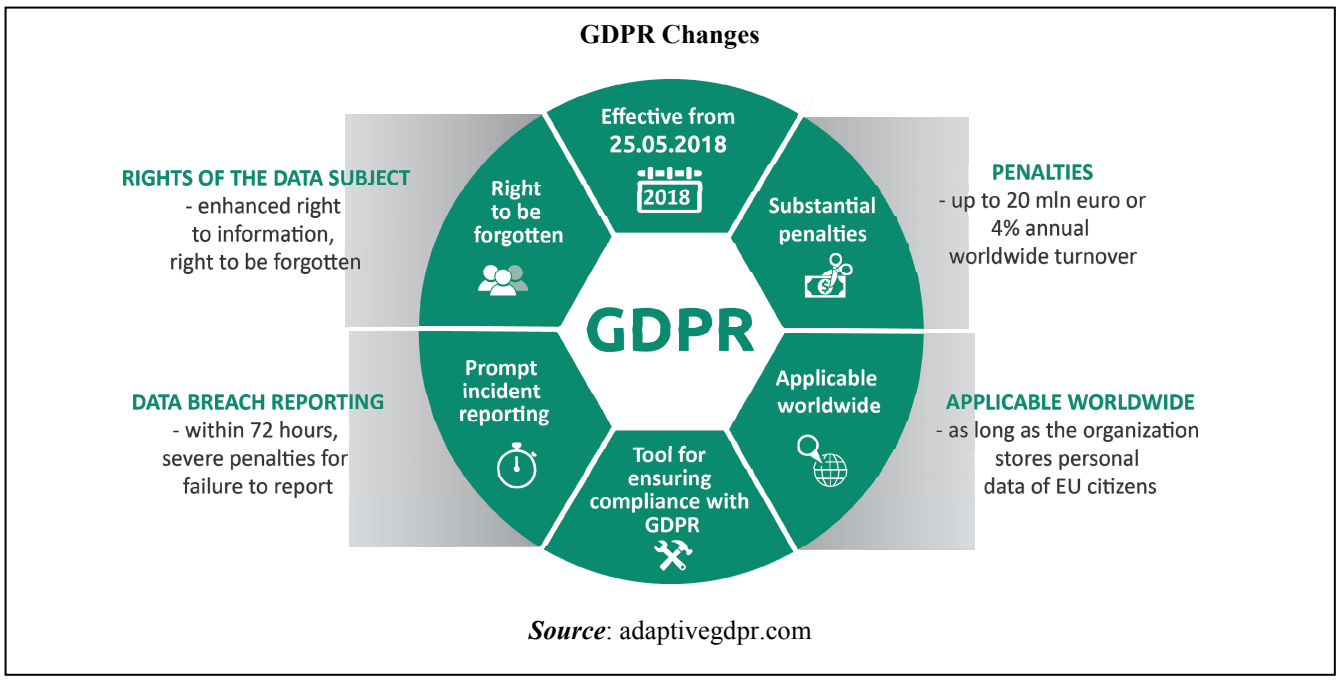
A. Data collection, repurposing

The data collection is the beginning of the data processing thus the GDPR protects the data subjects' privacy strictly in this period. If an organization is collecting data from EU citizens, including its own employees, then the GDPR applies, even if the organization is outside the EU. In other words, the GDPR applies to all companies processing the EU citizens' personal data, regardless of the company's location.

The GDPR remains strict with the collection of personal data, and the legal grounds are similar to the Data Protection Directive (DPD). The main legal basis is still the data subject's consent. The rules of collection and use of special categories of personal data (e.g., race, religion) also follow the language of the DPD, but the exemption for statistical and scientific research purposes for the processing of special categories of personal data is an important step toward the collection and use of personal data for scientific research. Furthermore, it is often not possible to fully identify the purpose of processing for scientific research at the time of the collection of personal data, and the GDPR recognizes this issue in Recital 33.

The purpose limitation principle means that personal data collected for one purpose should not be used for a new, incompatible purpose. This principle is maintained by the GDPR for the processing of personal data. When the personal data is processed for a secondary purpose, the DPD required that it could not be "incompatible" with the original one, while the GDPR requires the new purpose to be "compatible" with the original one. For the first blink, the negative approach of the DPD (the new purpose cannot be "incompatible"[2]) seems broader than the positive requirement in the GDPR for "compatible" repurposing. However, the drafting history shows that the repurposing was an important issue during the whole period.

The GDPR provides a list of factors to be considered for evaluating the compatibility with the new purpose, which is a step forward for acknowledging the new era of data processing. The boundaries are still not clear, and the member states have much space left to decide how strict they would like to be towards big data and the repurposing for scientific research.



Source: adaptivegdpr.com

In the EU law, a recital is part of the text, usually the beginning of the law, which explains the reasons for the provisions, and it is not normative, thus legally not binding. Recitals are usually general statements. The Recital of the GDPR gives guidelines for understanding the normative text and the purposes behind it. The Recital is more committed to the innovation, and open to the repurposing of processing than the normative text of the GDPR. One of the most important examples is the Recital 50, which states if the new data processing purpose is compatible with the original one, no separate legal basis is required. Furthermore, the statistical and scientific research purposes should be considered compatible lawful processing operations. It is questionable why this strong statement was not involved in the normative text. One of the earlier drafts of the GDPR contained an additional legal ground for the "statistical and scientific purposes" [3]. However, this provision was omitted from the normal text and mentioned just in the Recital part of the GDPR, leaving room to navigate for the member states. At the beginning of the drafting process, the Council may have intended to enable big data analysis for scientific research as a special legal ground.

The GDPR never mentions "big data", but in the final version, Recital 147 acknowledges its merits and methods especially in the health-related research by mentioning that "By coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions… On the basis of registries, research results can be enhanced, as they draw on a larger population.". The notion of coupling information for research is a step towards big data which clearly shows the evolution of the GDPR towards this type of data analysis and technological developments.

The researchers may also avoid restrictions on the repurposing of the processing of health data. The final text of the GDPR offers an exemption for the further processing for scientific research and statistical purposes at the purpose specification principle (Art. 5. (b)), similar to the DPD. The GDPR has been equipped with a broader definition of research (Recital 159), involving both public and private researching activities. The health-related research is also involved in this paragraph, which is important for the understanding the scope of the scientific research exceptions involved in the normative text. After the earlier drafts, the final version leaves no doubt that both private and public biomedical research activities are covered by the exemptions of the GDPR.

B. Notification

The data subject's right to be informed may raise a heavy burden for the researchers in the age of big data. Unlike the traditional research, where the purposes and goals were identified and communicated in the time of the data collection, the analysis of large datasets opened new horizons with the collected data, providing value by further processing and repurposing.

The period between the collecting and repurposing has a strong influence on the data controller's notification duties and the data subject's right to be informed. This gap may be bridged in many situations, but it seems impossible or needs disproportionate effort in many scenarios (e.g., the data subject changed her address, phone number, if they were provided to the data controller at all). The data controllers are required to provide updated notification of the new purpose to the data subjects as a general rule. However, the GDPR provides exemptions for the researchers from the notification requirement (Article 13 (5)). The Regulation also acknowledges that it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research (Recital 33).

Consenting to "certain areas" of scientific research means that the consent may be not required for a specific purpose, and researchers may have fewer notification duties on the uncertain future research in this specific area. This demonstrates that the GDPR permits more relaxed specificity

both with the notice and the consent provided for scientific research purpose. If the member states follow the Recital 33, the exemptions from the specific consent in the scientific research may be possible.

C. Profiling

Profiling is the automated processing of personal data to evaluate certain things about an individual, which can be part of an automated decision-making process. In other words, it can be a procedure which may involve a series of statistical deductions to make predictions about people. For instance, keeping a record of traffic violations to monitor driving habits of an individual to calculate his insurance fee. The GDPR restricts "profiling" by providing data subjects strong rights to avoid profiling-based decisions.

Restrictive rules on profiling are significant barriers for applying big data analysis in scientific research. If there is a decision based solely on automated processing which can have a legal or similarly significant effect on the individual, she has the right not to be subject to it. Further processing of personal data for scientific research was permissible under the DPD only if the member states furnished suitable safeguards that ruled out the use of the data in support of measures or decisions regarding the data subject. The Article 29 Working Party interpreted the 'measures or decisions' in the broadest sense, encompassing any relevant impact on particular individuals, either negative or positive. The broad interpretation of the Working Party led difficulties for the research projects applying automated processing. The GDPR is not as restrictive as the Directive by allowing further processing for research which can impact the individuals. As initially proposed by the Commission in 2012, the profiling provisions in the GDPR would have applied to any measure based solely on automated processing which "significantly affects" [4] the data subject. The vagueness of this wording could have unintended consequences for researchers using big data analysis since it is hard to define what kind of processing "significantly affects" [5] the data subject.

Collecting data from various sources (e.g., wearable devices, registries) and using it to evaluate the data subject's health conditions constitute "profiling" thus it would be important to clarify the requirements. The GDPR states that the data subject "shall have the right not to be subject to a decision based solely on automated processing, including profiling." This provision can be interpreted in two ways: as a prohibition or as a right to object. These interpretations can offer a different type of protection to the data subjects and the data controllers. If we interpret it as a prohibition, the data controllers would be obliged to meet with special conditions (e.g., explicit consent) to start the automated data processing which can significantly affect the data subject.The second interpretation, as a right to object, would let the data controllers start the processing, and the data subject has to act actively to protect her rights by restricting the automated processing [6].

D. Retention

There is a significant change in the length of the period when data may provide value for the researchers. In the past, the data was collected and processed for a specific research purpose to deliver the expected results, but in the age of big data, every piece of information has potential value. To put it another way, technological advancements and the new algorithms make the strict data retention periods questionable, especially in the field of scientific research.

Personal data may be retained as long as it is necessary for the purpose it was initially collected; thus, any data retention longer than that requires a new lawful purpose. Due to the purpose requirement, even a new consent alone is insufficient to extend the retention period, since the consent may be given for a specific purpose.

The GDPR also contains this traditional data retention principle (Recital 39, Article 5. (e)), similar to the DPD, but it opens toward big data with exemptions for statistical and scientific research (Recital 65, Article 5. (e)). Furthermore, the final version is less strict than the earlier drafts which required "periodic review to assess the necessity to continue the storage" as an additional condition for the longer retention period.

The laws regulating medical research and clinical trials usually require longer extended retention periods as a safeguard for the interest of the research subjects [7]. Thus, later the citizens can be approached if it is necessary (e.g., promising results of the new medicine), and the government authorities may also investigate the process and the results of the trial.

The extended retention periods are also tempting for the researchers to use the data for secondary purposes. Furthermore, de-identification techniques may mitigate the risks, thus the data can be harvested again.

E. Right to be forgotten

The right to be forgotten has emerged on the ground that the data subjects need more control over their personal data on the Internet. The search engines and the social networks changed the individual's privacy, and the answers from the EU were the new and stronger rights.

On the one hand, the "default of remembering" may negatively affect the individual's privacy. On the other hand, it may be beneficial for to society since it introduced the era of big data with the advanced computing systems and algorithms. Utilizing a large amount of health and other types of data has unleashed a huge wave of innovation, and the personalized healthcare services improved the quality of care. By making the "forgetting" default [8], or the right to be forgotten unconditional, there is a danger that privacy could prevail on the innovation.

The right to be forgotten applies to the processing of health data which means that the data subjects can request the erasure of their personal data from the researchers, but the GDPR provides exemption for scientific research purposes.

The definition of scientific research in the Recital is broad, including the privately funded research, which means if the member states follow the language of the Recital, it can protect the biomedical researching activities from the erasure requests of the data subjects, if it would render impossible or seriously impair the achievement of the objectives of the research. It can lead to harmonization issues that the placement of the definition of the scientific research in the Recital part of the GDPR may cause different interpretations of it by the member states.

The right to erasure can seriously hamper the biomedical research and affect the results of it. The drafters of the GDPR recognized this problem and tried to balance the researchers' and the data subjects' rights by offering a possible limitation of the above-mentioned rights in the case of scientific research as long as appropriate safeguards are implemented:

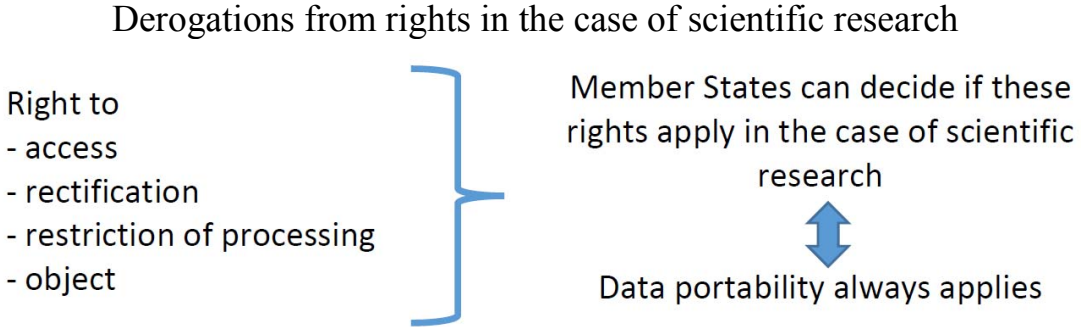
Article 17 (3) d) Paragraphs 1 and 2 (right to be forgotten) *shall not apply to the extent that processing is necessary: for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing;*

F. Right to data portability

The GDPR provide the data subject the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller, where the processing is based on a consent or on a contract and carried out by automated means (Article 20 (1)). The data portability was not just made to address the imbalance between the data controller and subject, but also to redress the economic imbalance between the large corporations, the biggest data controllers (e.g., Google, Yahoo). Therefore, this right has a serious impact on competition and consumer protection [9].
The data portability in the final version of the GDPR is significantly different, than in the earlier proposals: the format of the data, the object of the right, the balancing and the relationship with the right to be forgotten changed during the drafting period [10]. For the researchers, the most important novelties in the final version of the GDPR are the balancing clauses: the exercise of the right to data portability shall be "without prejudice" to the right to be forgotten (Article 20(3)) and "not adversely affect the rights and freedoms of others" (Article 20(4)).
However, compared to the right to be forgotten, the exemptions and balancing structure are completely different. Article 17(3) provides five specific cases in which right to erasure does not apply: freedom of expression; task carried out in the public interest or in the exercise of official authority; public interest in the area of public health; archiving, scientific or historical research purposes; and the establishment, exercise or defense of legal claims, while there are no references to the milder balancing clauses that we find in Article 20 ("without prejudice" and "not adversely affecting").

Article 20 (3) The exercise of the right referred to in paragraph 1 of this Article (right to data portability*) shall be without prejudice to Article 17.* (right to be forgotten*) That right shall not apply to processing necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.*
(4) The right referred to in paragraph 1 shall not adversely affect the rights and freedoms of others.

Furthermore, the GDPR Article 89 permits the Member States to provide derogations from the right to access, rectification, restriction, and objection of processing, when the personal data is processed for scientific research purposes. However, the data subjects retain their data portability rights, since the Member States cannot derogate from it [11]; thus, research institutes, companies and also healthcare providers have to comply with portability requests.
The application of portability rights may be an incentive for research, since the data subjects may transfer their data from one research institute to another, allowing other entities to conduct research with a wider database. Furthermore, the data subjects may make use of their data from other companies and smart devices. For instance, they can request the manufacturer of their smartphone, or smart watch to transfer their health and location data to a research institute which can use that data for scientific research or medical diagnosis.
On the other hand, this may create a burden for the researchers, since they must provide or transfer personal data, which was collected for research, and there is no clarity to the extent of the data that need to be transferred under a data portability request.



The data controller needs to transfer the personal data, which was 'provided' by the data subject. However, the meaning of the 'data provided' can be interpreted restrictively and extensively. If we interpret the scope of 'data provided' restrictively, it only involves the data which was provided explicitly by the data subject. For instance, in the beginning of the research, the data subject wrote down her name, basic health information, family history. According to the extensive approach, 'data provided' also involves data which was collected upon the consent or according to a contract (e.g., results of the research, the data collected during medical diagnosis after the consent).

G. Appropriate Safeguards

Appropriate safeguards have to be implemented by the member states for the application of the statistical and scientific exemptions of the GDPR, but the safeguards are not detailed in the GDPR, which makes the Regulation both technology neutral and future-proof, but it fails to harmonize the requirements on the EU level. The GDPR gives the member states the flexibility to draft their regulatory framework for big data and scientific research, even though the GDPR lays down the key principles and values. The specification of the safeguards is typically provided in legislation by the member states, which can be precise or more general, which leaves room for professional codes of conduct and/or further guidance released by the competent data protection authorities [12].

Based on the Data Protection Directive, the national data protection laws have different requirements for de-identification in the member states. Where the relevant ethical and legal standards are unclear, there is the potential for unnecessary regulatory delay for medical research projects [13].

The necessary level of de-identification is crucial among the safeguards which has to be harmonized on the EU level and the first proposals of the GDPR failed to do that. Recognizing the broad spectrum of de-identification and regulating certain key points by acknowledging as "appropriate safeguard" enables the development of regulatory guidance that encourages the maximum use of de-identification and it can open the way for using Big Data in scientific research. The European data protection law has taken a binary approach to the de-identification of personal data which means data is either personal data and therefore subject to data protection law, or it is anonymous thus not subject of the data protection law.

This binary approach can lead weaker level of data protection measures. When personal data is processed for a research purpose that cannot be accomplished with anonymized data, the researchers may have less incentive to use de-identification techniques, if they cannot reach any potential benefit by the data protection law (e.g., less strict rules would apply). Even if some level of de-identification would be compatible with the purposes of the research and could provide meaningful privacy protections for the individuals, there is no "reward" for the researchers' efforts by applying them, thus the binary approach of de-identification can result to weaker data protection [14].

The GDPR defines personal data as "any information relating to an identified or identifiable natural person." In the age of Big Data, more distinction should be made between the "identified" and "identifiable" personal data to recognize the appropriate level of safeguards and to use the potential value of the information, especially in the field of scientific research. Anonymization would be the strongest de-identification tool and safeguard which could solve the data protection issues of the Big Data use in research by excluding the application of the GDPR, but in many cases it cannot be applied because of the loss of data utility and value. Unlike anonymous data, pseudonymous data remains the subject of the Regulation and traditionally used techniques to protect privacy in research environment, such as key-coding, fall within the definition of pseudonymisation and therefore remains in the scope of the GDPR [15].

The final text of the GDPR highly promotes the pseudonymisation, which is the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information. Pseudonymization is not required in every research but rather the use of it is encouraged "as long as (the research purposes) can be fulfilled in this manner". Pseudonymization can offer a middle way between the directly identifiable personal data and completely anonymized data. By recognizing the value and importance of this de-identification method, the GDPR accepts that such data is potentially valuable and under less strict conditions can be reused [16].

By fostering pseudonymisation, the GDPR takes a big step toward the Big Data use in the field of scientific research and makes it easier for the member states to draft laws and accepting research activities without the requirement of anonymization. Pseudonymization's rise can be seen as a response to the Big Data and technological developments. The Article 29 Working Party in its opinion on the EU data protection reform proposals also argued that the pseudonymisation should be introduced more explicitly in the GDPR, as it can help to achieve a higher level of data protection [17]. The arguments of the Working Party showed that the Data Protection Officers of the member states highly promote the use of a middle-way de-identification tool, and it can predict the future interpretation of the GDPR by the member states, because the Data Protection Officers have significant impact on it in their countries and in the European Data Protection Board.

III. Conclusion

One of the main purposes of the GDPR was to protect the citizens' privacy in the online environment. However, the new and stronger rights, such as right to be forgotten and data portability may seriously hamper the scientific research and technological developments. It became clear during the drafting period that a proper balance needs to be found, otherwise the EU can fall behind in research and technological innovation. Thus, the GDPR introduced exemptions from almost every privacy protection rule, to enable the scientific research.

The main purpose of the GDPR was to unify the data protection rules in the EU, but it did not do it perfectly, since

many exemptions and important definitions are in the Recital part of the GDPR, which means it is not binding the member states. This situation may lead to different interpretations and forum shopping, which may erode the individuals' privacy, since the member states do not want to fall back in the field scientific research and losing income.

REFERENCES

[1] Viktor Mayer-Schönberger & Yann Padova, 'Regime Change? Enabling Big Data through Europe's New Data Protection Regulation' (2016) 17 Colum. Sci. & Tech. L. Rev. 315, 1.

[2] The Article 29 Working Party, 'Opinion 03/2013 on purpose limitation' 39.

[3] No. 9565/15 of 11 June 2015 "Processing of personal data which is necessary for archiving purposes in the public interest, or for historical, statistical or scientific purposes shall be lawful subject also to the conditions and safeguards referred to in Article 83."

[4] Brussels, 25.1.2012 COM(2012) 11 final 2012/0011 (COD) Proposal Article 20

[5] Article 29 Data Protection Working Party, 'Opinion 01/2012 on the data protection reform proposal' 14.

[6] Wachter, Sandra and Mittelstadt, Brent and Floridi, Luciano, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2016) SSRN: https://ssrn.com/abstract=2903469

[7] Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC Art. 58

[8] Mayer-Schönberger, Viktor, 'Delete: The Virtue of Forgetting in the Digital Age' (2011) Princeton University Press,

[9] Article 29 Data Protection Working Party, 'Guidelines on the right to data portability' (2016)

[10] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, Ignacio Sanchez, 'The right to data portability in the GDPR: Towards user-centric interoperability of digital services' (2018) Computer Law & Security Review, Volume 34, Issue 2, 193-203, ISSN 0267-3649

[11] Vayena E, Blasimme A, 'Biomedical Big Data: New Models of Control Over Access, Use and Governance' (2017) J Bioeth Inq. 508.

[12] Article 29 Data Protection Working Party, 'Opinion 03/2013 on purpose limitation' 28.

[13] Rumbold, John M. M., and Barbara K. Pierscionek. 2017. "A Critique Of The Regulation Of Data Science In Healthcare Research In The European Union". BMC Medical Ethics 18 (1). doi:10.1186/s12910-017-0184-y; 2.

[14] Mike Hintze, 'Viewing the GDPR Through a De-Identification Lens: A Tool for Clarification, Compliance, and Consistency.' (2017) 2.

[15] Gabe Maldoff. "How GDPR changes the rules for research." https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/. April 19, 2016. Accessed January 15, 2017.

[16] Viktor Mayer-Schönberger & Yann Padova, 329.

[17] Article 29 Data Protection Working party, 'Opinion 01/2012 on the data protection reform proposals'