Anonymised Data and the Rule of Law

*Daniel Groos and Evert-Ben van Veen**

The scope of application of the GDPR is determined by whether data are personal data or not, hence are anonymous data. By still insisting on Opinion 5/2014 the EDPB ignores that in 2016 the CJEU gave a different test to decide whether data are anonymous or not. Our proposal with the six safes test builds on that decision and will also bring the rule of law back in another essential dimension, namely legal certainty. The factors which decide whether data are anonymous or not can be influenced by the holder of the data, while Opinion 5/2014 states that anonymous data can become personal data again because of amongst other things new statistical techniques.

I. Introduction

Any Act needs a clear scope of application. If that is unclear, one might be subjected to the Act and even be fined for non-compliance, while one thought in good faith to remain clear of it. In the case of the General Data Protection Regulation (GDPR)[1] the primary scope of application is 'personal data', as defined in article 4.1 of the GDPR.[2] Formally the material scope is formulated in article 2 GDPR. In addition there is the territorial scope as defined in article 3.1 GDPR but we will not discuss those topics. If data are not personal data, the GDPR does not apply to anyone or anywhere. As will be discussed later, the European Data Protection Board (EDPB) favours a wide interpretation of personal data and still refers to[3] the Opinion 5/2014 on anonymisation techniques[4] of its predecessor, the article 29 Working Party under article 30.3 of Directive 95/46/EC, the Data protection directive (DPD).[5]

We will argue that Opinion 5/2014 has become obsolete after the Breyer decision of the European Court of Justice (ECJ).[6] We will also argue that the DPD definition of personal data did not materially change with the advent of the GDPR. Hence Breyer is still valid.

As almost all CJEU decisions the judgment is based on the issue at hand. We will forward a more generic description of the conditions when personal data are not personal data anymore according to Breyer based on 'the Five Safes' model,[7] which we extend to six safes, namely also the data in transit. Following our proposal will bring back the 'rule of law'[8] to this core aspect of the GDPR, both in a formal sense, being that ultimately a Court decides about the interpretation of an Act and not the regulator, and in a

* Daniel Groos, MLCF, the Netherlands. Evert-Ben van Veen, MLCF, the Netherlands. For correspondence: < eb.vanveen@mlcf.eu >. Research for this paper was partially funded by two H2020 projects to which we contribute, RECAP-preterm (grant number 733280) and HEAP-Exposome (grant number 874662). Colleague Martin Boeckhout helped to sharpen the focus of this paper during an early discussion of its outline. Matti Rookus of the Netherlands Cancer Institute provided valuable support and the input for the reference on SNP's. The two reviewers helped to improve the argument even though we might not agree completely. Their impartial contributions are greatly appreciated.

1 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

2 GDPR, Article 4.1.

3 At note 76 in the *Guidelines 05/2020 on consent under Regulation 2016/679*, version 1.1 adopted on 4 May 2020.

4 Article 29 Working Party (A29 WP), 'Opinion 05/2014 on Anonymisation Techniques' (10 April 2014) WP 216.

5 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data

6 Case C-582/14, *Patrick Breyer v Bundesrepublik Deutschland*, [2016], ECLI: EU: C: 2016:779.

7 Tanvi Desai, Felix Ritchie and Richard Welpton, 'Five Safes: Designing data access for research' (2016) < http://csrm.cass.anu.edu.au/sites/default/files/rsss/Ritchie_5safes.pdf > accessed 6 July 2020.

8 Martin Krygier, 'Rule of Law', in Michel Rosenfeld, András Sajó (eds), *The Oxford Handbook of Comparative Constitutional Law* (Oxford University Press 2012) 233-250; L. Pech, 'The Rule of Law as a Constitutional Principle of the European Union', (2009) 4 Jean Monnet Working Paper Series < http://jeanmonnetprogram.org/wp-content/uploads/2014/12/090401.pdf > accessed 17 April 2020. The latter especially stressing the 'Rechtsstaat' or 'Etat de droit' meaning that administrative decisions should be based on Acts and that the scope of application of such Act should be predictable.

more substantive sense, being that the law should not have a scope of application which is infinite and can be arbitrarily executed. Opinion 5/2014 and the Dutch Data Protection Authority in its wake,[9] state that because of 'new techniques', which the holder of the data cannot influence, one can never be sure whether anonymous data might become personal data again. Such an approach erodes legal certainty which is an essential aspect of the rule of law.[10] An Act which leaves the prime scope of application dependent upon unknown circumstances which one cannot influence, would not even be law in the Fullerian[11] sense. Our proposal explains how the threshold between personal data and anonymous data is guided by factors within the range of competence of all parties involved in the chain of data from the original controller to the holder of the then anonymised data.

We apply our proposal to the exchange of health data for scientific research. The stakes are high. The data exchange with research institutions in various research fields[12] [13] in the USA has stalled, largely because of the extensive interpretation of personal data. This GDPR interpretation of personal data has been called a 'sea change'.[14] It also hampers data sharing in other areas such as data exchange within the European Economic Area (EEA) given different legal bases for further processing of data for research or an initial consent not foreseeing the exchange of possibly still identifiable data with other researchers. Hitherto they were considered anonymous data. When further use of personal data for research is based on the 'consent or anonymise' approach[15] with often a smaller or broader public interest exception,[16] the anonymise route has de facto been made illusory because of the EDPB interpretation of anonymous data.

Anonymisation certainly does not mean that from then on everything is allowed. Anonymisation is a form of data processing and should not only have a legal basis but should also have a sound ethical basis to then further process the anonymised data. Not everything goes once data have been anonymised. We will briefly discuss that aspect but refer to the work of others for principles which we subsume under 'good health research governance'.

Much of what is discussed in this paper builds on the work of others. During the research for this paper and working on the 'Five Safes' model, the new book of Arbuckle and El Emam appeared which helped to connect the dots.[17] Also in other papers, to be cited later, we found that our intuitions were not so original as we had thought and had been expressed by others. In the conclusion we will come back to the fact that the EDPB seems to ignore this body of literature.

II. Opinion 5/2014

In 2014, during the height of the debate around the draft GDPR, the article 29 WP issued the Guidelines on anonymisation techniques.[18] The thrust of the guidelines is as follows. It should not be possible:
- to single out an individual;
- to link records relating to an individual, or

9 The Report published on 16 December 2019 with the findings of the investigation into the data processing of SBG relies heavily on Opinion 5/2014. The decision in the complaint procedure against SBG published on the same data states that it uses the Breyer criterion, yet relies heavily on the non-Breyer criteria employed in the investigation. < https://www.autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/rapport_bevindingen_sbg_en_akwa_ggz.pdf > < https://www.autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/beslissing_op_bezwaar_sbg.pdf > accessed 6 July 2020.

10 See(n 8).

11 Lon L. Fuller, *The Morality of Law* (revised edn, Yale University Press 1969).

12 Tania Rabesandratana, 'European data law is impeding studies on diabetes and Alzheimer's, researchers warn' (Science, 20 November 2019) < https://www.sciencemag.org/news/2019/11/european-data-law-impeding-studies-diabetes-and-alzheimer-s-researchers-warn > accessed 6 juli 2020.

13 For cancer research; personal communication of Matti Rookus

14 David Peloquin, Michael DiMaio, Barbara Bierer & Mark Barnes 'Disruptive and unavoidable: GDPR challenges to secondary research uses of data' (2020) 28 European Journal of Human genetics 697.

15 Nayha Sethi, Graeme T Laurie, 'Delivering proportionate governance in the era of eHealth: Making linkage and privacy work together' (2013) 13 Med law Int 168.

16 G. Owen Schaefer, Graeme Laurie, Sumytra Menon, Alastair V. Campbell & Teck Chuan Voo, 'Clarifying how to deploy the public interest criterion in consent waivers for health data and tissue research', (2020) 21 BMC Medical Ethics < https://bmcmedethics.biomedcentral.com/track/pdf/10.1186/s12910-020-00467-5 > accessed 6 July 2020.

17 Luk Arbuckle, Khaled El Emam, *Building an Anonymization Pipeline: Creating Safe Data* (O'Reilly Media 2020).

18 Opinion 05/2014 on Anonymisation Techniques, (10 April 2014).

– to infer information about an individual.

Though the Opinion mentions 'means likely reasonably to be used',[19] it states that all anonymisation should be completely irreversible, warning against new technologies which could make datasets earlier presumed anonymous, re-identifiable after all. [20]

> Opinion 5/2014 even goes so far as to state:
> "Thus, it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data".[21]

In their very critical appraisal of the Opinion El Emam and Alvarez[22] challenge this statement.[23] If the statement would be taken literally, it would mean that all Open Data which statistical agencies publish, would still be personal data as of course the underlying microdata[24] cannot be deleted. This also applies to Open Data of governmental agencies under the Directive 2013/37/EU[25] and to the statistically relevant outcomes of research while the underlying granular research data remain unchanged and should remain unchanged as outcomes of empirical research need, apart from the FAIR principles,[26] to be reproducible.[27] Nobody in their right mind would consider the highly aggregated data which we read in the news or scientific papers personal data because the data on which they are based have not been deleted at event level. [28]

The EDPB still refers to this Opinion in its recent Guidelines on consent of May 2020.[29] A few weeks earlier the EDPB stated in its COVID-19 Guidelines[30] that anonymised data means that it is no longer possible for *anyone* (our emphasis) to refer back to the original data subjects.[31] The Dutch government referred to Opinion 5/2014 in a letter to Parliament about further use of health data for research.[32]

However, it can be seriously doubted whether Opinion 5/2014 reflects good law. As will be discussed below, the CJEU had a more nuanced vision and the final text of the GDPR did not include 'singling out' as a criterion for considering data personal data but only as an example which could render data more easily identifiable.[33]

III. The CJEU in Breyer

1. Introduction

In October 2016 the CJEU issued its decision in the Breyer case. At issue in Breyer was mainly whether a dynamic email address should be considered personal data in the German circumstances. The conclusion was that in this case a dynamic IP address should be considered personal data. Yet, much more important is the test which the Advocate General and the CJEU

19 As stated in recital 26 of Directive 95/46/EC, (n 5).

20 Opinion 5/2014, (n 4)

21 ibid 9.

22 Khaled El Emam, Cecilia Álvarez, 'A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques' (2015) 5(1) International Data Privacy Law < https://doi.org/10.1093/idpl/ipu033 > accessed 6 juli 2020.

23 For a critique, see also Michele Flinck, Frank Pallas, 'They who must not be identified- distinguishing personal data form non-personal data under the GDPR' (2020) 10 International Data Privacy Law 11.

24 About microdata of statistical agencies see amongst others, OECD Expert Group for International Collaboration on Microdata Access, final report, OECD, Paris, July 2014.

25 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information.

26 Mark D. Wilkinson et al, 'FAIR Guiding Principles for scientific data management and stewardship' (2016) 3 Scientific Data < https://www-nature-com.eur.idm.oclc.org/articles/sdata201618.pdf > accessed 6 July 2020.

27 Steven N. Goodman, Danielle Fanelli, John O.A, Ioannidis, 'What does research reproducibility mean?' (2016) 341 *S*cience Translational Medicine 341.

28 See for the transformation in such data before they can be published as open data: Thijs Benschop, Matthew Welch, 'Statistical Disclosure Control for Microdata: A Practice Guide for sdcMicro' (2016) < https://sdcpractice.readthedocs.io/en/latest/ > accessed 6 July 2020.

29 At note 76, EDPB Guidelines 05/2020 on consent under Regulation 2016/679, version 1, adopted on 4 may 2020. In the section on research these Guidelines nearly literally repeat the Guidelines on consent of the 29 WP (article 29 WP Guidelines on Consent under Regulation 2016/679 (wp259rev.01) of June 2018 .

30 Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, 21 April 2020.

31 Ibid at note 17.

32 Kamerstukken 2019-2020, 27529, nr. 191, 3. However, more recently, in the context of sharing data of mobile phones to find clusters of people meeting, the government referred to Breyer, Kamerstukken 2019-220 35479, nr. 3 at 6.

33 GDPR, Recital 26. Nevertheless, we agree that singling out in the online environment via tracking cookies, even if the internet user is not identifiable in the GDPR sense, raises important privacy concerns. In Europe that issue is addressed in Directive 2002/58/EC, the present ePrivacy Directive.

employed. That test differed substantially from the 5/2014 Guidelines.

It should first of all be noted that all the deliberations of both the AG and the Court could have cut short if simply just 'singling out' would be sufficient to consider data also personal data. A dynamic IP address obviously singles out, at least in some point of time, but that was not sufficient for the AG or the Court to consider that address personal data.

2. The Advocate General (AG)

Amongst other things the role of the AG is to reflect on the literature and possibly relevant soft law. Sometimes the AG refers concurrently to the article 29 WP or the EDPB Opinions or Guidelines.[34] That did not happen in this case. On the contrary. Not only did AG Campos Sanchez not refer to Opinion 5/2014 in her Opinion,[35] she also criticised rather explicitly an earlier Opinion of the art 29 WP on the concept of personal data.[36]

The AG stresses in short two points when discussing 'means likely reasonably to be used by the controller or by any other person'. [37] First, 'by any person' should not be seen as any conceivable third party. That 'overly strict interpretation' would never rule out with absolute certainty that a third party would be capable of revealing a person's identity'.[38] Second, 'means likely reasonably to be used' does not mean any means, but reasonable and not prohibited means.[39] With both points the Opinion chooses for the so called 'relative approach'. Not whether it is in theory possible to link, but whether it is in reality possible to reidentify by the controller with the legitimate help of a known third party. The AG explicitly did not want to expand the concept of identifiability beyond those situations as it would lead to legal uncertainty when data would ever cease to be personal data.

3. The CJEU

The Court leaves the question of relative or absolute approach in the middle but referring to the AG the CJEU agrees on the test itself. The combination of two known parties matters.[40] It then states:

> This, as the Advocate General stated essentially in point 68 of his Opinion, that (means likely reasonably to be used for identification, authors) would not be the case if the identification of the data subject was prohibited by law or practically impossible on the account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant.[41]

In that sense the CJEU is more inclined to the relative approach, as is also shown in point 49 of the decision.

The concrete outcome of the judgment was that in the German situation the website holder would have legitimate means to identify the data subject via referring to the internet provider which assigned the IP address in the case of a cyber-attack to the website.

IV. Puzzling about Breyer and a Possible Way out of the Puzzle

Some comments to the decision read in it what they always believed to be true, namely that also Breyer hardly leaves room for anonymised data.[42] If the CJEU had opted for the relative approach, a dynamic IP address couldn't be categorized as personal data in relation to a website publisher, since the provider lacks the information that would be needed to identify Breyer without a disproportionate ef-

34 See for instance Case C-673/17, *Planet49*, [2019], ECLI:EU:C:2019:801, Opinion of Advocate General Szpunar, which references Opinion 04/2012 (par. 40), Opinion 2/2010 (par. 81), Opinion 15/2011 (par. 81), Opinion 2/2010 (par. 108), Opinion 2/2013 (par. 108), Opinion 15/2011 (par. 118); Case C-40/17, *Fashion ID*, [2019], ECLI:EU:C:2019:629.

35 Case C-582/14, Patrick Breyer v Bundesrepublik Deutschland, [2016], ECLI: EU: C: 2016:779, Opinion of Advocate General Campos Sanchez-Bordona, delivered on 12 May 2016.

36 Article 29 Working Party (A29 WP), 'Opinion 4/2007 on the concept of personal data', (20 June 2007) WP 136.

37 Recital 26 of the GDPR uses the same wording but reversed reasonably and likely.

38 At point 65.

39 At point 68.

40 At point 45, also at point 48.

41 At point 46.

42 Frederik Zuiderveen Borgesius, 'The Breyer Case of the Court of Justice of the European Union: IP Addresses and the Personal Data Definition' (2017) 1 EDPL 130; Frank van 't Geloof, 'CJEU: Dynamic IP Addresses as Personal Data', (2017) 1 CRi 26; P. Quinn, L. Quinn, 'Big genetic data and its big data protection challenges' (2018) 34 Computer Law & Security Review 1000.

fort. Therefore, 'reasonably likely' should be understood as to mean 'absolutely impossible'.[43]

Yet most others have come to a different conclusion.[44] In sum: it is relevant *who* has access to the information needed to identify the data subject and the CJEU is taking steps towards a risk-related approach. Furthermore, a study commissioned by the European Parliament, Panel for the Future of Science and Technology (hereinafter referred to as the STOA study)[45] notices a difference between Opinion 5/2014 and Breyer, describing the Breyer test as being 'more pragmatic'.[46]

We concur with the latter type of comments. In essence the CJEU moved away from a zero risk interpretation of personal data. Opinion 5/2014 refers to abstract statistical techniques for anonymisation. Only if those techniques are being followed to the full, and nobody could possibly reidentify, the data can be considered anonymous.

Breyer on the other hand requires a concrete test for the data at hand, hence the result, and the context in which the data are being processed. Actually Breyer gives us two distinct tests:

1. For a controller who is not prohibited by law to identify, the data would be anonymous if the identification requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant.
2. For a controller who does not meet the first test, so the risk of identification is in reality not insignificant, the data would still be anonymous if identification either by that controller or with the help of a known third party was prohibited by law.

This does not mean that Breyer does not leave us with certain puzzles. The first test is clear from a legal point of view but needs to be operationalised in practice, as will be explained in section 7.

The second test is also troublesome from a legal point of view.[47] A hacker could gain access to these data. Hacking is an illegitimate act in almost all jurisdictions[48] but that does not mean that it might not happen.

We should add an additional criterion to that test, in line with test 1, being:

– and that the risk of identification by a third party using illegitimate means would require a disproportionate effort in terms of time, cost and manpower, so that that risk of identification appears in reality to be insignificant.

In this context it should be remembered that the data under the second test were already not directly identifiable.

V. Relation of the Ruling to the Definition of Personal Data Under the GDPR

As mentioned, Opinion 5/2014 was issued during the debate about the new GDPR. The EP wanted 'singling out' as part of the definition of personal data[49] and the EDPB may have wanted to support the EP in this proposal but the result of the trialogue was that singling out did not become part of the definition. The political outcome of the trialogue was not to broaden the scope of personal data as compared to Directive 95/46/EC[50] The definition of personal data in ar-

43 Van 't Geloof 2017, 27.

44 Miranda Mourby et al, 'Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK', (2018) 34(2) Computer Law & Security Review 222; R. P. Santifort, 'Naar een meer genuanceerde benadering van 'pseudonimisering in het privacyrecht', (2019) 5 Privacy & Informatie 195; K. Demetzou, 'Data Protection Impact Assessment: A tool for accountability and the unclarified concept of 'high risk' in the General Data Protection Regulation' (2019) 35(6) Computer Law & Security Review < https://www-sciencedirect-com.eur.idm.oclc.org/science/article/pii/S0267364918304357 > accessed 6 July 2020; A. El Khoury, 'Personal Data, Algorithms and Profiling in the EU: Overcoming the Binary Notion of Personal Data through Quantum Mechanics', (2018) 3 Erasmus Law Review 165.

45 European Parliament, How the General Data Protection Regulation changes the rules for scientific research, Study, Panel for the Future of Science and Technology, ERPS, European Parliamentary research service, scientific foresight unit (STOA) PE 634.447, July 2019, < https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634447/EPRS_STU(2019)634447_EN.pdf > accessed 6 juli 2020.

46 ibid 30.

47 See Mark Phillips, Edward S. Dove & Bartha M. Knoppers, 'Criminal Prohibition of Wrongful Re-identification: Legal Solution or Minefield for Big Data?' (2017) 14 Bioethical Inquiry 527.

48 See Philippe Jougleux, Lilian Mitrou, Tatiana Eleni Synodinou 'Criminalization of Attacks Against Information Systems', in Ioannis Iglezakis (ed.), *The Legal Regulation of Cyber Attacks* (Kluwer Law International, 2020); see also Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union.

49 Albrecht, draft Report 17 December 2012 (COM (2012)001).

50 The in the Netherlands authoritative "Tekst en Commentaar' states' that the GDPR did not mean to broaden the concept of personal data and that 'singling out' was not taken up as a criterion to consider data personal data. Zwenne, G.J., Knol, R.C., (eds), (*Privacy- en telecommunicatierecht*, Wolters Kluwer, 2018) 70.

ticle 4.1 of the GDPR gives more examples of identification but did not change either.

In the new Recital 26 'singling out' remained only as one of the criteria which can make data more easily identifiable. The Recital still uses the same phrase for the (re) identification test but reversed the terms 'likely' and 'reasonably', hence it became 'means reasonably likely to be used'. This change resulted in better English but is materially insignificant.

There is also another change in Recital 26 DPD versus Recital 26 GDPR. Recital 26 DPD stated "means likely reasonably to be used (...) either by the controller or by *any* other person to identify...." (our emphasis). Recital 26 GDPR states "means reasonably likely to be used (...) by the controller or *another* person to identify (...)" (our emphasis). Mourby considers this change a support for the relative approach as we have seen in Breyer.[51] Whether that is true or not, it is certainly not a support for the absolute approach.

Hence, Breyer, issued under mentioned Directive, reflects the definition of personal data of the GDPR as well.

However, the GDPR introduced pseudonymisation, being:

> the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.[52]

This does not change the conclusion above. As also follows from Recital 29, the separation between identifiers and the pseudonym (P) relates to that procedure at the controller. At the controller P is reversible, subject to safety measures. This has sparked the debate whether pseudonymised data which arrive at another controller, are still personal data (assuming that the data under P are not personal data, see hereinafter) if the new controller does not have the legal or reasonably practical means to reverse P into the identity of data subject. With Mourby[53] and Santifort[54] we can concur they could then be anonymous data, however, only after following the test explained in section 7.

A different discussion is about "pseudonymised data" when P is *not* reversible, such as generated by a secure one-way hash. That is not pseudonymisation in the sense of the GDPR,[55] even though we are usually referring to those data as pseudonymised data as well. A new term would be helpful.[56] Whether those data are personal data depends on the robustness of how P is generated and the data under P, in particular when data from various data sources under the P arrive at a new data holder. We will operationalise that test in section 7.

VI. The Rule of Law: Relation Between Article 29 WP /EDPB Soft Law and the CJEU

Opinions, recommendations and the like of administrative agencies are examples of 'soft law'. Soft law is a common phenomenon in the regulatory state, also in the European Union. [57] Soft law is not binding.[58] It can increase legal certainty as the regulator clarifies how it interprets the legislation.

Soft law cannot be challenged directly in a court. It can be challenged via 2 routes.

The relatively direct route is that, when a decision of an administrative agency is based on soft law, the court will first of all hold that soft law against the light of the background legislation and consider whether it is a proper translation of that background legislation.[59]

The Breyer case is an example of the most indirect route. The court is asked to give an opinion about the law and in the background the soft law interpreta-

51 Miranda Mourby, 'Anonymity in EU health law: not an alternative to information governance' (2020) 0 Medical Law Review 1, 11.

52 Article 4.5 GDPR

53 Mourby, (n 44).

54 Santifort, (n 44).

55 European Union Agency for Cyber Security (ENISA), Pseudonymisation techniques and best practices, Recommendations on shaping technology according to data protection and privacy provisions, November 2019, exemplifies that in the case of pseudonymised data in sense of the sense of the GDPR, 'by definition' there should be a recovery mechanism, from the 'pseudonymisation secret' to identity of the data subject, 25.

56 Evert-Ben van Veen, 'Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate', (2018) 104 European Journal of Cancer 70.

57 Chalmers, D., Davies, G., Monti, G., *European Union law*, Cambridge University Press, 2019, 116-119.

58 Article 288 Treaty on the Functioning of the European Union.

59 See for a strking example in the case of competition law College van Beroep voor het bedrijfsleven, 17-03-2020, ECLI:NL:CBB:2020:177

tion of the background legislation plays a role. In the Breyer case that was not Opinion 5/2014 but the earlier and also 'expansive' interpretation of personal data in Opinion 4/2007. That interpretation was explicitly refuted by the AG and the CJEU came to a different test as explained above.

While admitting the various interpretations of the rule of law under legal scholars, one of its pillars is that in the end the court decides and not an administrative agency.[60] In that sense it is somewhat disappointing the EDPB never reconsidered its Opinion 5/2014 in the light of this decision, assuming that it could ignore the criticism in the literature.[61] The EDPB referred to Breyer only on one occasion, namely that dynamic IP addresses are personal data.[62] That is not what Breyer actually decided. The CJEU ruled that *in the German situation* (our emphasis) dynamic IP addresses were personal data. And it is certainly not the main point of Breyer. In other matters the EDPB quite often refers to CJEU decisions,[63] and even updated a previous opinion in light of a new decision.[64]

VII. Towards Operationalisation of the Breyer Tests

There seem to be several reasons why 5/2014 still is considered the norm instead of the Breyer test. As seen, the EDPB still promotes it. Many Data Protection Officers, which all research institutions must have instituted,[65] are not always lawyers and will be inclined to follow the EDPB guidelines and the like to the letter and may not be interested to delve into more nuanced case law. The third reason may be that the Breyer test seems to be less clear than the absolute approach of Opinion 5/2014. As, how to operationalise 'in reality insignificant'? The zero chance approach of Opinion 5/2014 means to err on the extremely safe side, though as we have seen in the introduction, at the expense of data exchange or with the consent or anonymise approach, at the expense of the usability for the intended purposes. There is always a trade-off between the usability of the data for research and the level of anonymisation as explained in the ISO document on privacy enhancing data de-identification terminology. [66]

Data in health research needs to be sufficiently nuanced to lead to valid conclusions. Research, especially in genomics, needs to relate to large numbers of participants.[67] Somewhat paradoxically, in order to achieve 'personalised medicine', larger sets of data must be examined to find sufficient statistical valid correlations for those smaller subgroups.[68] There will be in terms of Arbuckle and El Emam[69] 'permutations' applied to the data in order to assure data security and, as it would be called under the GDPR, 'privacy by design and default'[70] [71] which will eliminate direct identifiability. However, there is always a trade-off between the usability of the data for research and the level of anonymisation.[72] That trade-off should be established by the methodological re-

60 See (n 8).

61 See (n 23 , 24, 52).

62 Article 29 Working Party (A29 WP), 'Opinion 01/2017 on the Proposed Regulation for the ePrivacy Regulation (2002/58/EC)', (4 April 2017) WP 247, 27.

63 See for instance Opinion 28/2018 regarding the EU Commission Draft Implementing Decision on the adequate protection of personal data in Japan, which references CJEU cases C-362/14, C-203/15, C-293/12 and C-594/12, pp. 10, 22, 25; Opinion 23/2018 on Commission proposals on European Production and Preservation Orders for electronic evidence in criminal matters, which references CJEU decisions C-203/15 en C-698/15, 12,14.

64 Update of Opinion 8/2010, WP 179 Update, in light of CJEU decision C-131/12.

65 Article 37 GDPR.

66 see ISO/IEC 20889:2018(en) Privacy enhancing data de-identification terminology and classification of techniques, < https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en > accessed 6 July 2020.

67 See eg Lee, A. et al, 'BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors', (2019) 21 Genetics in Medicine 1708; Mavaddat, N. et al, 'Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes', (2019) 104 The American Journal of Human Genetics 21.

68 See for a critical approach with further references: Klaus Hoeyer, 'Data as promise: Reconfiguring Danish public health through personalised medicine', (2019) 49(4) Social Studies of Science 531.

69 Luk Arbuckle, Khaled El Emam, Building an Anonymization Pipeline: Creating Safe Data (O'Reilly Media 2020).

70 Article 25 GDPR.

71 As early as 2008, see the, as it was then called PET (privacy enhancing technologies) described in: Evert-Ben van Veen, 'Obstacles to European research projects with data and tissue: solutions and further challenges', (2008) 44 European Journal of Cancer 1438; and amongst many others Kuchinke W, Ohmann C, Verheij RA, et al, 'A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model' (2014) 83(12) International Journal of Medical Informatics < https://pubmed-ncbi-nlm-nih-gov.eur.idm.oclc.org/25241154/ > accessed 6 July 2020.

72 see ISO/IEC 20889:2018(en) Privacy enhancing data de-identification terminology and classification of techniques, < https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en > accessed 6 July 2020.

quirements of the research, which means that in health research these data need to be sufficiently granular. The analyses made on those data aim for statistically relevant correlations about properties of a group of patients based on the chosen indicators/parameters (e.g. certain SNP's,[73] lifestyle, environmental factors, treatment) and outcomes such as disability, social functioning, occurrence of disease or death. Each of the participants needs to be individually discerned by a random number. As argued, if that number is generated by a one-way hash, that would not make the data pseudonymised in the GDPR sense.

The question is when in that data chain can data be considered anonymous. Breyer calls for a contextual approach. Such an approach has been made concrete by Arbuckle and El Emam in their recent book, with the following initially rather easy formula: the risk of re-identification is dependent upon: the content of the data and the context in which these data are processed.[74] They provide various instances of the combination of transformations of the data and the contexts in which they can be processed which will decrease the risks of reidentification.

When the context is more open, the data should become less granular. The most open context is 'open data' as published by statistical agencies and governmental bodies. There are no limitations or checks how such data will be used. Hence these data must meet the highest level of anonymisation. Statistical agencies have developed standards for this.[75] Here we are interested in data which are used in intermediary research stages where various statistical analyses will be employed on the data before the results are published.

Those steps are the context. The safety of the context can be described via the 'Five Safes' model,[76] in short:[77]

1. Safe projects: is this use of the data appropriate?
2. Safe people: can the researchers be trusted to use it in an appropriate manner?
3. Safe data: is there a disclosure risk in the data itself?
4. Safe settings: does the access facility limit unauthorised use?
5. Safe outputs: are the statistical results non-disclosive?

As there is in these projects always a chain of data, we should add a sixth element: safe transport. It is not only about data in situ but also about data in transit; can the data not be intercepted during transit. But admittedly that would be one of the easiest ones in the row of 6; namely are the data sufficiently encrypted during transit. Technically this can usually be solved by a yes or no answer, while the other elements of 'safe' require an evaluation and the sum of the various elements decides whether the data may be used for anonymisation and can be considered anonymous.

We will discuss the safe projects in the next section. Here we are interested in elements 2-4.

1. Ad element 2

Desai et al.[78] and El Emam[79] give examples of assuring safety for the second element. I have once suggested that a researcher does not have any real interest in re-identification. It would mean the end of his or her career.[80] In the Netherlands all researchers are subjected to the Code of Conduct on research integrity,[81] in other countries similar safeguards will exist. But the possibility of a researcher going astray and becoming an inside adversary can never be fully excluded. This setting can never be set on completely safe.

2. Ad element 3

This chance of reidentification via those data is remote even though smart statisticians have shown that this is sometimes and in some cases possible for

73 See e.g.: Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium, Breast Cancer Association Consortium, (2006) 99(5) J Natl Cancer Inst.

74 Luk Arbuckle, Khaled El Emam, Building an Anonymization Pipeline: Creating Safe Data (O'Reilly Media 2020) 52.

75 Steven N. Goodman, Danielle Fanelli, John O.A, Ioannidis, 'What does research reproducibility mean?' (2016) 341 Science Translational Medicine 341.

76 *Desai* (n 7).

77 ibid 5.

78 *Desai* (n 7).

79 See(n 17).

80 E.B. van Veen, 'Europe and tissue research: a regulatory patchwork', (2013) 19(9) Diagnostic Histopathology 331.

81 Netherlands Code of Conduct for Research Integrity 2018 < https://www.knaw.nl/shared/resources/actueel/bestanden/netherlands-code-of-conduct-for-research-integrity-2018-uk > accessed 6 July 2020.

genetic data.[82] Those are abstract cases where actually the statistical researchers making those few linkages with identifiable persons were acting illegally when processing personal data without a legal basis. Outside adversaries are not likely to be interested in datasets which need advanced statistical techniques to decipher and never have mail addresses or passwords attached which can be used on the black market of spamming or even extorsion.[83]

3. Ad element 4

Yet, the main safeguards are in our opinion to be found in element 4. The safest case would be a research surrounding where one can do all the analyses but can only export the statistical outcomes. This does not need to be a kind of bunker without access to the internet as described by Mourby et al.[84] One might work in another department with access to this safe digital research environment or even from home, given that all cyber security threats are being averted.[85] Statistics Netherlands has such a 'remote access' facility.[86] The microdata at Statistics Netherlands are obviously not anonymous data to Statistics Netherlands. They can be considered anonymous data to the researchers when they bring their own data to be matched with the data Statistics Netherlands, as there is no possibility to retrieve the identity of the data subjects in the statistical outcomes which are allowed to leave the analysis platform.

Yet, this approach might lead to new national data silo's and will not always work for 'data lakes' in combined research efforts. Given the other safes, the data transfer agreements (DTA's) where assurances will be made how and in which safe data environment that data will be processed, also the receiving research institution can be considered to receive anonymous data as it cannot reasonably retrieve the data subject. Those databases should be ISO 27001/2 certified with control via logging and other safeguards controlling the use of the data. Together with the other safes and accountable assurances in the DTA's, the data could still be considered anonymous.

VIII. Bringing Back Legal Certainty

One of the advantages of this proposal is that it brings back legal certainty. With the expansionist vision one never knows whether one is data controller or not. Opinion 5/2014 stresses that data might be anonymous now but could be personal data in the even near future either because of replay back of the hashing mechanism which generated the code number or by new statistical techniques for matching while scraping the internet where people might have added new data about themselves. It was also an argument of the Dutch AP in the SBG case.[87]

With any Act there will be borderline cases where there may be reasonable doubt whether a certain situation falls under its remit or not. But an Act where the regulator pushes the scope of application into the unknown because of 'new techniques', fails to do what any Act should in a liberal democracy, namely bringing legal certainty. That was basically one of the arguments of the AG in Breyer, which regretfully seems to be overlooked in the present debate.

Our proposal for the contextual approach with the six safes also leaves a margin for which side of the threshold certain data fall. But in this case the data holder can influence on which side the data will fall. New threats to the safe environment can make data identifiable after all, also in the contextual approach. The data holder must remain alert and apply state of the art techniques by which it can be ascertained that those threats are avoided in practice. Further linking by which the combined data would fall into a different class of possible identifiability according to the third safe element, would also require an action of the data holder. Obviously, those measures of the data holder should be demonstrable and auditable. Nonetheless, one does not suddenly go from data holder to data controller with all the responsibilities attached to it, simply because of new theoretical techniques or because someone decided to put all his or her data on the internet.

82 Yaniv Erlich, Tal Shor, Itsik Pe'er, Shai Carmi, 'Identity inference of genomic data using long-range familial searches' (2018) 362 Science 690.

83 The argument for this aspect of the risk based approach was made by the ICO in: Anonymisation: managing data protection risk code of practice, November 2012, available at: https://ico.org.uk/media/1061/anonymisation-code.pdf

84 Mourby, (n 44).

85 See for example the Citrix vulnerability in 2019: https://support.citrix.com/article/CTX267027

86 https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research, accessed 7 July 2020.

87 Note 9.

IX. This is Not About Escaping the GDPR: Towards Good Research Governance

As seen, the Opinion 4/2007 promoted an extensive interpretation of the concept of personal data. Opinion 4/2007 also mentioned that data protection legislation would have sufficient nuances to accommodate for various levels of possible reidentification,[88] with presumably different regimes. The Opinion did not give examples of such differentiation. They hardly exist at the moment either. The GDPR could give some leeway not to notify the data subjects after a data breach if the data have been properly pseudonymised.[89] Article 11 GDPR solves the paradox that if proper privacy design has been employed and the data controller cannot reach the data subjects as it is lacking direct identifiers, the controller would not breach the GDPR because of not notifying the data subjects or not giving them rights of access etc. For such controllers articles 15 to 20 GDPR do not apply unless the data subject would submit additional information by which the controller can retrieve the data subject in the database. In some EU member states regulations have been put in place, implementing 9.2.i and 89.1 GDPR, that further use of patient data for research is allowed without consent if these data have been pseudonymised.[90] Other member states have similar exceptions where often further conditions will apply, such as the ethical vetting of research.[91]

However, also with the clearer cut off point the GDPR is not out of scope in the data chain. Anonymised research data starts with once personal data. Anonymisation is a form of data processing and should be compatible with the original purpose. The last part of article 5.1b GDPR states that further processing for statistic or research purposes is not incompatible with the original purpose if the conditions of article 89.1 are being met. Yet, with the EDPS,[92] we do not see this as a freeway for any research, even when the data has been anonymised. If consent for research purposes was the original legal basis, the following research should not be incompatible with that original consent, whether the data have been anonymised or not. If consent for research was not the original legal basis of the data processing, as will be the case in most jurisdictions for data processing in the interaction between patients and their physicians, anonymization for research should still be research which can claim to have social value[93] and meet the requirements for good research as one of us has stated elsewhere.[94] The research should meet the reasonable expectations of the data subjects, a phrase from Recital 50 pertaining to article 6.4. We do not state here that both 5.1.b and 6.4 should be met. That does not make sense from a dogmatic point of view as in that case 5.1.b might just as well not have been written.

The solution is that there are requirements to the research which may legitimately use 5.1.b. There is quite some literature about the views of patients about further use of patient data and they do not always come down to informed consent in the sense of the GDPR.[95] According to Skovgaard et al, that seems more an obsession of the researchers which set up the empirical research.[96]

88 *Patrick Breyer v Bundesrepublik Deutschland* (n 36) 25.

89 Art. 34.3.a GDPR mentions that the communication with the data subject in case of a personal data breach is not required if the controller has implemented appropriate measures that "*render the personal data unintelligible to any person who is not authorised to access it.*" In addition to pseudonymisation, also robust encryption will play a role here.

90 Johan Hansen, Petra Wilson, Eline Verhoeven, Madelon Kroneman, Mary Kirwan, Robert Verheij, Evert-Ben van Veen, Assessment of the EU Member States' rules on health data in the light of GDPR, Report commissioned by the European Commission in the context of the Third EU Health Programme, December 2020, pending publication.

91 ibid.

92 European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, at p. 22 and following.

93 See Shona Kalkman et al, 'Responsible data sharing in international health research: a systematic review of principles and norms' (2019) 20 BMC Medical Ethics <https://bmcmedethics.biomedcentral.com/track/pdf/10.1186/s12910-019-0359-9 > accessed 6 July 2020.

94 Evert-Ben van Veen, 'Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate' (2018) 104 European Journal of Cancer 70.

95 E.g. Coppen R, van Veen E-B., Groenewegen P.P., Hazes J.M., de Jong J.D., Kievit J., de Neeling J.N., Reijneveld S.A., Verheij R.A., Vroom E. (2015) 'Will the trilogue on the EU Data Protection Regulation recognise the importance of health research?' (2015) 25(5) Eur J Public Health 757; Gesine Richter, Christoph Borzikowsky, Wolfgang Lieb, Stefan Schreiber, Michael Krawczak, Alena Buyx, 'Patient views on research use of clinical data without consent: Legal, but also acceptable?' (2019) 27 European Journal of Human Genetics 841; Skovgaard L, Wadmann S, Hoeyer K, 'A review of attitudes towards the reuse of health data among people in the European Union: The primacy of purpose and the common good', (2019) 123 Health Policy 564.

96 Skovgaard, note 92.

We subsume social value, transparency, meeting reasonable expectations and accountability under good research governance, meaning not how research is governed by state actors, as it is often understood,[97] but how it governs itself in the light of its social responsibilities in interaction with the main stakeholders, patients and the public at large. The need for such 'good research governance' is exacerbated by the fact that the results of research apply to everybody belonging to the group to whom those results (might) relate, whether you have indirectly contributed to those results as a data subject or not.

A principled approach to data in the whole research data chain is needed.[98] It would require a separate paper to delve deeper in this self- or co-governance, we may refer to valuable contributions elsewhere.[99] These developments in co-governance can be seen as a refinement of the tradition of ethical vetting to which health research is subjected for decades in almost all Western countries.[100] With large multi-centre observational research projects there will be multiple control. The lack of alignment of the research ethic boards is problematic.[101] Not generally the lack of oversight, even if Data Protection Authorities would not be involved because of the anonymised nature of the data.

X. Concluding Remarks

We have proposed an approach to personal data which will bring the rule of law back in the discussion. Both in the sense of taking the decision of the CJEU about the concept seriously and by defining the circumstances by which a data holder can organise on which side of the threshold it will fall, personal data or anonymous data. We might go a step further and wonder whether, when still insisting on Opinion 5/2014 and downplaying Breyer and ignoring the vast body of literature which critiques that Opinion, the EDPB acts in true constitutional spirit. Though the EDPB tells us to stand for our fundamental rights on data protection, it should be aware that there are more fundamental and more persistent aspects of our constitutional ordering as well, being the rule of law as we explained earlier.

Recently the EDPS referred to both Breyer and Opinion 5/2014.[102] That was a step forward but still ignores that the Opinion and Breyer were incommensurable in their approaches and incompatible in the outcomes of the respective tests. We fully agree with the EDPS that research exemptions, such as 5.1b GDPR last sentence, should only be applied to, as the EDPS calls that, 'genuine research'. Ohers have used the phrase 'bona fide' research.[103] Section 9 of this paper discussed the criteria for such research. Section 7, about the six safes criteria, shows that accountability, a fundamental aspect of the GDPR, also applies to the data holder in the data chain as to why data can be considered anonymous. Both safeguards taken together, anonymisation is not a way to escape the GDPR and underlying values completely. But it is a way to get data exchanged again and safely used for legitimate purposes without undue hindrances.

97 As in most contributions to the Oxford Handbook of Governance, D. Levi-Faur (ed.), *The Oxford Handbook of Governance* (Oxford University Press 2012).

98 Bart van der Sloot, *Privacy as Virtue. Moving beyond the individual in the age of big data* (1st edn, Intersentia 2017), 148-156.

99 Kieran C O'Doherty et al, 'From consent to institutions: designing adaptive governance for genomic biobanks' (2011) 73 Soc Sci Med 367; Bartha Maria Knoppers, 'Framework for responsible sharing of genomic and health-related data' (2014) The HUGO Journal (2014) 8 Hugo J < https://thehugojournal.springeropen.com/articles/10.1186/s11568-014-0003-1 > accessed 6 July 2020; Lea NC, Nicholls J, Dobbs C, et al, 'Data safe havens and trust: toward a common understanding of trusted research platforms for governing secure and ethical health research' (2016) 4 JMIR Med Inform < https://www-ncbi-nlm-nih-gov.eur.idm.oclc.org/pmc/articles/PMC4933798/ > accessed 6 July 2020.

100 See amongst others T.A. Faunce, *Pilgrims in Medicine: conscience, legalism and human rights* (1st edn Koninklijke Brill, Leiden, 2005) 160-183.

101 David Townend, Edward S. Dove, Dianne Nicol, Jasper Bovenberg, Bartha M. Knoppers, 'Streamlining ethical review of data intensive research' (2016) 354 British Medical Journal.

102 European Data Protection Supervisor (EDPS) 'Opinion 3/2020 on the European strategy for data' (16 June 2020).

103 See the discussion in Evert-Ben van Veen, 'Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate', (2018) 104 European Journal of Cancer 70.