scripted

Volume 18, Issue 1, September 2021

Biomedical Data Identifiability in Canada and the European Union: From Risk Qualification to Risk Quantification?

Alexander Bernier and Bartha Knoppers***

DOI: 10.2966/scrip.180121.4

Abstract

Data identifiability standards in Canada and the European Union rely on the same concepts to distinguish personal data from non-personal data. However, courts have interpreted the substantive content of such metrics divergently. Interpretive ambiguities can create challenges in determining whether data has been successfully anonymised in one jurisdiction, and whether it would also be considered anonymised in another. These difficulties arise from the law's assessment of re-identification risk in reliance on qualitative tests of 'serious risk' or 'reasonable likelihood' as subjectively appreciated by adjudicators. We propose the use of maximum re-identification risk thresholds and quantitative methodologies to assess data identifiability and data anonymisation relative to measurable standards. We propose that separate legislation be adopted to address data-related practices that do not relate to demonstrably identifiable data, such as algorithmic profiling. This would ensure that regulators do not expand the jurisprudential conception of identifiable data purposively to capture such practices.

Keywords
General Data Protection Regulation, data protection law, health data, information philosophy, privacy law, risk quantification

* Centre of Genomics and Policy, McGill University, Montreal, Canada, alexander.bernier@mail.mcgill.ca.
** Canada Research Chair in Law and Medicine, Director, Centre of Genomics and Policy, McGill University, Montreal, Canada, bartha.knoppers@mcgill.ca.

The authors extend their gracious thanks to Mark Phillips for providing extensive comments and suggestions on multiple drafts of this manuscript.

1 Introduction

The health sector is highly reliant on data to perform scientific research and to guide administrative and policy decision-making. Data privacy and data protection legislation in Canada and the European Union regulate the use of identifiable personal data. Individuals enjoy substantive rights in their personal data. Considerable procedural preconditions are also imposed on the use and disclosure of such data by regulated entities. Health entities therefore place high reliance on anonymised data in performing collaborative international uses of information whilst respecting the privacy rights and data protection interests of individuals. Open science initiatives, for instance, often grant researchers and the public liberal access to anonymised data. Population health and public health initiatives utilise aggregate or summary-level data from multiple countries or environments for the purpose of deriving comparative insights. Therefore, clarity as to the threshold of identifiable data is instrumental to the efficient exchange of health data amongst entities in the health sector. If the distinction between personal data and anonymised data is ambiguous or non-harmonised, healthcare institutions and scientists may refrain from the collaborative use of anonymised data for fear of non-compliance with restrictions on the use of personal data.

The legislative approach which frames data identifiability in qualitative language that requires regulators and the judiciary to subjectively assess the residual risk of individual re-identification renders ambiguous the scope of application of data privacy and data protection law. Such indeterminacy exposes regulated entities to significant risk of legal non-compliance. This may induce health institutions to limit their secondary use of health information to avoid the risk of legal non-compliance. It may further limit the effectiveness of best-efforts legal compliance, as the indeterminacy of legal identifiability criteria could make

it difficult to differentiate identifiable regulated data from non-identifiable unregulated data even in cooperation with seasoned legal counsel.

We contend that it would be preferable to use rigorously defined quantitative language and risk-modelling methodologies to express concepts of data identifiability. This would bolster the predictability of data privacy and data protection law, promoting heightened data use in the health sector.

We also propose that certain interests in data, including risks arising from algorithmic profiling and data mining, that are not reliant on the use of data that is demonstrably identifiable, should be addressed using a different legislative framework than privacy or data protection. Our argument is structured as follows.

First, we demonstrate how divergent conceptions of 're-identification risk', 'personal identifier', and 'personal relation', distinguish identifiability metrics in Canada and the European Union. Our analysis concludes that there is significant overlap, but not total harmonisation, between the legal identifiability metrics applicable in Canada and the European Union.[1]

Second, we consider how potential threats to individual privacy, and certain intended regulatory targets of data privacy legislation, including algorithmic decision-making and individual profiling, do not require the use of demonstrably identifiable personal data. Our approach recognises that 'anonymised' data can create sensitive probabilistic inferences about specified

[1] See also: Mark Phillips and Bartha Knoppers, (2016) 34 *Nature Biotechnology* 1102-1103. Phillips and Knoppers provide a parallel analysis of the non-harmonization of health data identifiability metrics in the health sector.

individuals, even without unauthorised access to data that bears a clear relation to such persons.[2]

Drawing from informatics literature and information philosophy, we demonstrate certain limitations of conventional qualitative data identifiability metrics in addressing both privacy and other values enshrined in data privacy and data protection legislation. This analysis further explains the mutable nature of data identifiability standards. Courts may broaden the scope of identifiable data to ensure that penalties and remedies can be applied to potentially objectionable data uses regardless of whether such practices truly implicate the use of identifiable data as conventionally understood.

The potential for this to happen arises as "identifiable data" plays a double role in data privacy and data protection legislation. Individuals must demonstrate that their identifiable personal data has been processed as a precondition actualizing their substantive rights. Identifiable data also functions as an "object" relative to which data users (i.e., controllers and processors) must discharge procedural duties in processing all instances.[3]

Therefore, a narrow definition of identifiable data can limit the substantive rights of individuals concerning data use practices. Conversely, a broad definition of identifiable data can significantly increase the procedural burden and compliance risk that regulated entities face. To remedy these difficulties, we propose the use of separate legislation to capture harmful data-related practices that do not require the use of demonstrably identifiable personal data.

---

[2] Jeffrey Skopek, "Big Data's Epistemology and Its Implications for Precision Medicine and Privacy" in I Cohen and others (eds.), *Big Data, Health Law, and Bioethics* (CUP: 2018), pp. 36-41. Worku Urgessa, "The Protective Capacity of the Criterion of "Identifiability" Under EU Data Protection Law" (2016) 2 *European Data Protection Law Review* 521-531.

[3] Dara Hallinan and Raphaël Gellert, "The Concept of 'Information': An Invisible Problem in the GDPR" (2020) 17:2 *SCRIPTed* 269.

The adoption of maximum re-identification risk thresholds for data to be considered non-identifiable and defined de-identification methodologies to produce anonymised data is proposed to clearly distinguish personal regulated data from non-personal unregulated data. These measures can ensure that data identifiability receives a consistent interpretation, without leaving stated regulatory targets of data privacy laws such as algorithmic decision-making and profiling unregulated where no demonstrably identifiable data is used to perform such actions.

For greater clarity, the distinctions between "qualitative" approaches and "quantitative" approaches to assessing data identifiability can generally be described as follows. Qualitative approaches require humans to exercise their judgment to subjectively determine if there exists a serious risk of data being re-identified, within the context of its use. As shall further be discussed, quantitative approaches formalise the assumptions made in assessing the residual risk of re-identification inherent in data using statistical measures and mathematical models. That is, quantitative approaches express data identifiability as a residual risk score calculated relative to formalised assumptions. Our contention is not that quantitative metrics of data are infallible, but rather that the rigorous definitions afforded thereby better allow for the comparison of assumptions and expectations regarding residual identifiability than do qualitative metrics. This affords greater certitude to the law.

Third, we perform a survey of emergent quantitative methodologies that are being used to calculate and compare data identifiability and measure privacy harms, especially in the health sector. We conclude that regulators could adopt such quantitative metrics to encourage the meaningful comparison of data identifiability in different contexts against predetermined benchmarks. Such approaches can further be helpful in formalising the content of privacy interests in an internally consistent manner, which can be difficult to achieve using

traditional statutory language. We now turn to the comparative analysis of data identifiability standards in Canadian law and European Union law.

2 Comparing Legal Standards of Data Identifiability

2.1 Is it identifiable or anonymous?

The first feature assessed is whether a reasonably foreseeable prospect of identification exists, within the factual context of the data's use. Identifiability is used in Canada and the European Union to distinguish regulated data from unregulated data.

2.2 Canada's dual requirement: (a) "personal" information "about" an (b) "identifiable" person"

Canada's federal and provincial legislation, case law, and regulatory guidance inform data identifiability standards. Privacy statutes use inconsistent definitions of identifiable personal data across jurisdictions (provincial and federal) and sectors (private, public, health).[4] Considering only statutory definitions, Canada's data identifiability regime appears non-harmonised. Some laws utilise a test of reasonably foreseeable prospect of identification alone or in combination with other data, others advance a more restrictive definition considering "identifiable" only data which readily re-identifies its subject, and still others leave the term undefined.[5] Courts and regulators have generally harmonised

[4] Noela Inions, Leanne Tran, and Lorne Rozovsky, *Canadian Health Information: A Practical Legal and Risk Management Guide* (LexisNexis, 2018), pp. 19, 22-24.

[5] Council of the Canadian Academies, "Accessing Health and Health-Related Data in Canada" (2015), p. 195.

identifiability standards, converging on equivalent[6] tests of "serious possibility"[7] or "reasonable expectation"[8] of identification as the inclusion threshold of identifiable data.[9]

Case law considers data personal if it satisfies a relational or substantive "personal" metric and a contextual "identifiable" metric.[10] First, an analysis of the data's content must reveal its personal character.[11] Second, a "serious possibility" of individual identification must inhere in the data.[12]

One must not conflate the "personal" nature of data and the "personal" nature of the identifier. Ascribing data a "personal" character means inquiring if it is "private", or is "about" or sufficiently "related to" the concerned identifier. Ascribing the "identifier" a personal character requires assessing if the 'object of identification' to which the data bears a "serious possibility" of being associated demonstrates a sufficient connection to an individual. The former consideration is addressed here.

---

6 *Canada (Information Commissioner) v Canada (Public Safety and Emergency Preparedness)* [2019] FC 1279 (hereinafter Safety), paras 52-54.

7 *Gordon v Canada (Minister of Health)* 2008 FC 258, (hereinafter *Gordon*), para 34.

8 *Canada (Information Commissioner) v Canadian Transportation Accident Investigation & Safety Board* [2006] F.C.J. No. 704, (hereinafter *NavCan*), paras 9, 49.

9 Office of the Privacy Commissioner of Canada, "Real Fears, Real Solutions: A Plan for Restoring Confidence in Canada's Privacy Regime" (2017), p. 26.

10 *Dagg v Canada (Minister of Finance)* [1997] S.C.J No. 63, (hereinafter *Dagg*), para 68; *Gordon*, *supra* n. 7, paras 34-43.

11 *University of Alberta v Alberta (Information and Privacy Commissioner)* 2009 ABQB 112, (hereinafter *Alberta-University*), paras. 63-68.

12 Normann Witzleb and Julian Wagner, "When Is Personal Data about or Relating to an Individual a Comparison of Australian, Canadian, and EU Data Protection and Privacy Laws" (2018) 4 *Canadian Journal of Comparative and Contemporary Law* 293-330, pp. 310-315.

2.2.1 Canada's "personal" information "about" criterion

Case law and investigation reports posit three approaches[13] to delimiting "personal data." The narrowest approach requires "personal data" to demonstrate a privacy interest.[14] The broadest approach deems all data "relating to" an identifiable individual "personal" data.[15] Privacy as consecrated in Canada's civil rights legislation and constitutional laws informs the narrow approach. The narrow approach interprets privacy statutes purposively and reads in limits to personal data, borrowing from Canadian criminal law jurisprudence that weighs the individual privacy interest against the societal interest in public knowledge. This approach further relies on the narrow

---

[13] For greater clarity, there is little or no disagreement in Canadian case law as to the legal test to be used to determine if data constitutes identifiable personal information. *Dagg* and *NavCan* establish the consensus legal test used to determine if data is personal. *Gordon* establishes the consensus legal test to determine if data is identifiable. The different approaches described above relate to conflicting jurisprudential approaches to the interpretation and application of the consensus legal tests.

[14] *Husky Oil Operations Ltd. v Canada-Newfoundland and Labrador Offshore Petroleum Board* 2018 FCA 10, (hereinafter *Husky-Oil*), paras 23-31, 45-46; *Suncor Energy Inc. v Canada-Newfoundland and Labrador Offshore Petroleum Board* 2018 FCA 11, paras. 17-18; *Leon's Furniture v Alberta (Information and Privacy Commissioner)* 2011 ABCA 94, (hereinafter *Leon*), paras. 47-51; *Otis Canada Inc. v I.U.E.C., Local 1* [2010] B.C.W.L.D. 7681, (hereinafter *Otis*), paras. 66-77, 88-92; *Board of Education of School District No. 68 (Nanaimo-Ladysmith)* [2008] B.C.I.P.D. No. 28, (hereinafter *Nanaimo-Ladysmith*), para. 50; *NavCan*, *supra* n. 8, paras 47-54; *Hamilton Police Services Board, Re* 2006 CarswellOnt 11827, paras. 22-23, 42; *Ontario (Ministry of the Solicitor General), Re* [2003] O.I.P.C. No. 281, (hereinafter *Solicitor*), paras. 32-39, 45; *Toronto (City), Re* 1993 CarswellOnt 7190, para. 9.

[15] *Zelstoff Celgar Ltd* [2017] B.C.C.A.A.A. No. 28 (hereinafter *Zelstoff*), paras. 48-51; *Janssen-Ortho Inc. v Canada (Minister of Health)* 2007 FCA 252, paras. 8-9; *Schindler Elevator Corp., Re* [2012] B.C.I.P.C.D. No. 25, (hereinafter *Schindler*), paras. 82-85; *Calgary Police Service, Re* [2013] A.W.L.D. 906, paras 8-9; Information and Privacy Commissioner of Ontario, *Ministry of Community Safety and Correctional Services* (2009), at 7-9; *Canada (Information Commissioner) v Royal Canadian Mounted Police Commissioner* 2003 SCC 8, paras 23-25; Information and Privacy Commissioner of Ontario, *Ministry of Health and Long-Term Care* (2001), (hereinafter *Health*), p. 6-8, confirmed in *Ontario (Attorney General) v Pascoe* [2002] O.J. No. 4300, paras. 2-7; Information and Privacy Commissioner of Ontario, 'The New Federated Privacy Impact Assessment (F-PIA): Building Privacy and Trust-Enabled Federation' (2009), p. 6.

conception of privacy as reflected more generally in Anglo-American tort law and constitutional law.[16]

The "broad" approach derives from international data protection standards, and especially the continental European conception thereof.[17] It employs a textualist reading of privacy statutes and relies on "identifiability" to limit the breadth of "identifiable personal data." The balance of competing rights (including privacy) is performed *after* data has been deemed to "relate to" an "identifiable" person, rather than in characterising "personal" data.[18] Such approach conceptualises data protection as imposing procedural requirements on data use generally, even absent a privacy interest. Privacy provides substantive guarantees forbidding proscribed data uses, that are not presumptively engaged by all data uses.[19]

The third, median approach posits "personal" data as a broad concept bounded by the statutory construction of "*personal*" data "*about*" an individual. Limiting factors include restricting "personal" data to categories analogous to those enumerated in non-exhaustive statutory definitions of "personal" data,[20]

---

16 *R. v. Tessling*, 2004 SCC 67, paras. 25, 59-62; Office of the Privacy Commissioner of Canada, "Privacy Law Reform: A Pathway to Respecting Rights and Restoring Trust in Government and the Digital Economy" (2019) (hereinafter OPC 2018-2019 Annual Report), pp. 11-21, 57; *Schindler*, *supra* n. 15, para 68. Congressional Research Service, "Data Protection Law: An Overview" (2019). Paul M. Schwartz and Karl-Nikolaus Peifer, "Transatlantic Data Privacy Law" (2017) 106 *Georgetown Law Journal* 115-180.

17 Menno Mostert and others, "From Privacy to Data Protection in the EU: Implications for Big Data Health Research" (2018) 25 *European Journal of Health Law* 43-55. Nadezhda Purtova, "The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law" (2018) 10 *Law, Innovation, and Technology* 40-81, pp. 42-45. Raphaël Gellert and Serge Gutwirth, "The Legal Construction of Privacy and Data Protection" (2013) 29 *Computer Law and Security Review* 522-530.

18 *Prince Edward Island Health, Re* 2015 CarswellPEI 34, paras. 36-50.

19 Norberto de Andrade, "Data Protection, Privacy and Identity: Distinguishing Concepts and Articulating Rights" *Privacy and Identity Management for Life*, vol 352 (2011) 90-107, pp. 94-98.

20 *Ontario (Ministry of Government Services), Re* 2010 CarswellOnt 18874; paras. 7-13; *Alberta-University*, *supra* n. 11, paras. 71-72. *Sheptycki, Re* 2007 CarswellAlta 2695, paras. 17-21; *Ontario*

reliance on explicit statutory derogations from personal data (e.g. professional or public data), or interpreting data "*about*" an individual to require more than "mere reference" to the individual but less than a "privacy interest."[21] Multivalent approaches arise due to differing drafting and purposes[22] across Canada's many privacy statutes,[23] and to adjudicators' interpretive disagreement as to the meaning of "*personal*" data "*about*"[24] an identifiable[25] individual.[26]

(Ministry of Attorney General) Re, 2007 O.I.P.C. No. 179 (hereinafter *Attorney General*), paras. 38-43.

21 *Zelstoff*, *supra* n. 15, para. 48; *Alberta Health*, *Re* [2013] A.W.L.D. 412, (hereinafter *Alberta-Health*), paras. 41-48; *Ontario (Government Services), Re* 2012 CarswellOnt 1786, paras. 10-13; *Alberta-University*, *supra* n. 11, para. 72; *Ontario (Attorney General) v Fineberg* [1996] O.J. No. 67, paras. 4-5.

22 The statutory construction of Canadian data privacy law differs according to the sector the legislation pertains to (i.e., public, private, or health sector) and, in the federal public sector legislation, the statutory construction and legal reasoning applied by courts can further differ across the public data privacy context (i.e., relating to State collection, use, and disclosure of data) and the public access to information context (i.e., relating to the State's legal obligation to publicly disclosure data on request).

23 *Zelstoff*, *supra* n. 15, para. 48; *Schindler*, *supra* n. 15, paras. 82-85.

24 *Husky-Oil*, *supra* n. 14, paras. 31-46; *Nanaimo-Ladysmith*, *supra* n. 14, para. 50; *Safety*, *supra* n. 6, paras. 69-71.

25 *Dagg*, *supra* n. 10, paras. 58-87.

26 The disparate influence of continental European data protection law and Anglo-American privacy law on Canada's data privacy should be considered. Such influences may further explain the varying jurisprudential determinations in Canada's data privacy case law as to the role a privacy interest plays in delineating the boundaries of personal data. Canadian data privacy law reflects the influence of European data protection law, which conceptualises data protection as a legal regime that is distinct from privacy and is applicable to a broader ambit of data. This framing is established in EU legal instruments including the Charter of Fundamental Rights of the EU and the GDPR. Data protection rights are presumptively engaged by uses of personal data writ large and impose positive obligations on data controllers that are more expansive than the obligation of non-interference enshrined in the right to privacy. A broad concept of personal data is required to allow courts to assert jurisdiction on interferences with informational rights that do not directly impugn individual privacy, such as the right to ensure the accuracy of information or to be informed of automated decision-making. Canadian data privacy law is also influenced by the Anglo-American common law, which often relies on the limitative conceptions of privacy enshrined in constitutional law and tort law to restrictively interpret the content of data privacy statutes. Privacy is framed as a negative right to freedom from interference with a protected core of personal data that is private in nature, typically arising in specific contexts such as State surveillance or consumer transactions that create a heightened potential for the transgression of individual privacy interests, or a heightened expectation of confidentiality and privacy.

2.2.2 Canada's 'identifiable' criterion

Contextual identifiability considers whether the factual circumstances of data use create a reasonably foreseeable prospect of a natural person being identified. We favor the contextual approach, though it does not enable "absolute" data de-identification as does, for example, the Safe Harbor method of the United States' *Health Insurance Portability and Accountability Act* (HIPAA). HIPAA's absolutist approach fosters procedural certainty, efficiency, and the standardisation of de-identification methods. This is achieved by accepting a "one-size-fits-all" de-identification methodology to produce anonymised data,[27] unless the anonymising party has "actual knowledge" that the data remains identifiable.[28] The contextual approach determines identifiability from the circumstances of the data's use. Consequently, future changes thereto can alter data's legal identifiability. Nonetheless, the contextual approach more accurately reflects data's true re-identification risk.

Canada uses the contextual approach. Contextual determination of identifiability in Canadian jurisprudence assesses the "serious possibility" of identification inherent to the circumstances of data use. Data[29] and health data[30] de-identification guidelines, as well as limited case law,[31] describe relevant contextual factors. Incentives[32] for illicit data access, the deterrent value of

---

27 Mehmet Kayaalp, "Modes of De-identification" *AMIA Annual Symposium Proceedings 2017* (2018) 1044-1050.

28 "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule § 164.514(b)(2)(ii)" (Department of Health and Human Services, 2012).

29 Information and Privacy Commissioner of Ontario, "De-Identification Guidelines for Structured Data" (2016), p. 10.

30 Khaled El Emam et al., "Pan-Canadian De-Identification Guidelines for Personal Health Information" (2007).

31 *Alberta-University*, *supra* n. 11, paras. 64-65.

32 *Sa Majesté la Reine c. Rothmans Inc. et al*., 2019 NBBR 44 (hereinafter *Rothmans*), para. 141.

potential detection, the content and auditability of data-sharing agreements[33] and policies,[34] and the availability of the data are considered.[35] So are the likelihood of inadvertent identification, of internal misuse, and of external security breach.[36] Re-identification methods with an indeterminate potential for success are also relevant to the analysis. Re-identification risk is assessed relative to the practices of the data user,[37] or for contested disclosures, the anticipated practices of prospective data recipients.[38] Certain provincial regulators propose in regulatory guidance documents that the "serious possibility" of identification be calculated from[39] the likelihood that an event potentially causing identification occurs, multiplied by the probability of such occurrence causing re-identification.[40]

In sum, Canada's identifiability test first considers if data raises a privacy interest, or sometimes simply a relation to an identifier absent a privacy interest, and second, if the data presents a contextually arising "serious possibility" or "reasonable expectation" of individual identification.

2.3 The European Union's dual requirement: information (a) "relating to" (b) an "identifiable" person

The GDPR establishes the substance of EU data protection law, saving for Member State derogations. Our analysis is restricted to the general EU-applicable regime established by the GDPR and does not address potentially divergent Member State conceptions of personal data.

---

33 *Ibid* at para. 64; *PIPEDA Report of Findings No. 2015-001,* Re 2015 CarswellNat 868, para. 97.
34 *FortisBC Energy Utilities, Re* [2016] B.C.W.L.D. 3102 (hereinafter *FortisBC*), paras. 53-58, 84-85.
35 El Emam and others, "Pan-Canadian De-Identification Guidelines for Personal Health Information", *supra* n. 30, p. 52.
36 Information and Privacy Commissioner of Ontario, *supra* n. 29, p. 13.
37 *Rothmans*, *supra* n. 32, para. 141; *FortisBC*, *supra* n. 34, pp. 53-58, 84-85.
38 *Langley School District No. 35, Re*, [1998] B.C.I.P.C.D. No. 56, paras. 35-38.
39 *Vancouver Coastal Health Authority, Re*, [2003] B.C.I.P.C.D. No. 41, paras. 44-56, 68-72.
40 Information and Privacy Commissioner of Ontario, *supra* n. 29, pp. 11-16.

2.3.1 The European Union's 'relating to' criterion

The test enshrined in the GDPR[41] and in its predecessor the Data Protection Directive[42] reprises the "broad" Canadian approach to "personal data", including all data "relating to an identifiable natural person." The "relating to" criterion is broader in the GDPR than in Canada. In Canada, personal data "about" an individual must relate to them in content. The GDPR further incorporates data unrelated to the individual in content, the processing of which "is likely to"[43] affect the individual's interests in result, or which is used purposively to such effect. The GDPR considers personal phenomenal or environmental data not bearing on the individual, relating to them only in its context-specific analytic use.[44]

Relation exists if the data's "content" concerns the identified natural person, if the data is used for the 'purpose' of influencing them, or if the data will have the "result" of influencing the person's interests.[45] In *Nowak*, the Court of Justice of the European Union (CJEU) rejected any requirement for a privacy interest to inhere in personal data.[46]

In *Nowak*, a test examinee's answers and the examiner's comments were deemed the examinee's personal data.[47] Similarly, the subjective comments of

---

41 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, (hereinafter 'GDPR'), art. 4(1).

42 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, art. 2(a).

43 Article 29 Data Working Party, "Opinion 4/2007 on the Concept of Personal Data" (2007) 01248/07/EN WP 136, pp. 9-12.

44 Purtova, *supra* n. 17, pp. 52-55.

45 Article 29 Data Working Party, "Opinion 4/2007 on the concept of personal data", *supra* n. 43 at pp. 10-12.

46 *Peter Nowak*, Case C-434/16, [2017] EU:C:2017:582, [2017] OJ C 72, (hereinafter *Nowak*), paras. 34-35.

47 *Ibid*, paras. 37-42.

health professionals made in individuals' health records, and the inferences drawn about individuals from the genetic information of their biological relatives, are likely said individuals' personal data.[48]

The CJEU's prior holding in *YS* tempers *Nowak's* wide definition of "personal" data. *YS* establishes that an immigration officer's "[abstract] legal analysis" is not personal data, though the personal data restated in an analytical statement remains so.[49] By analogy, diagnostic reasoning likely constitutes personal data only insofar as it applies medical knowledge to a patient's data, to the exclusion of general medical principles restated in, but not relating to, patient records. The expansive European criteria for deeming data to be personal are justified by the wide range of data protection rights anticipated in the GDPR. Many such rights do not engage a privacy interest, but rather create broad rights to data protection in all instances of data processing that relate to or affect the concerned individual. To be actionable, such rights require data protection law to capture data generally having some connection to a person, rather than only a person's private data.[50]

2.3.2 The European Union's "identifiable" criterion

Breyer[51] establishes the principal EU identifiability test. Prior debate centered on whether identifiability is objective (relative to all parties) or relative (relative to

---

48 Daniel Jove, 'Peter Nowak v Data Protection Commissioner: Potential Aftermaths Regarding Subjective Annotations in Clinical Records' (2019) 5 *European Data Protection Law Review* 175-183, pp. 177-179.

49 *YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S*, Joined Cases C-141/12 and C-372/12, [2014] ECLI:EU:C:2014:2081 (hereinafter *YS*), paras. 13, 39-48.

50 Gellert and Gutwirth, *supra* n. 17; Mostert and others, *supra* n. 17.

51 *Patrick Breyer v Bundesrepublik Deutschland* Case C-582/14, [2016] EU:C:2016:779 (hereinafter *Breyer*).

the controller or processor).[52] According to *Breyer*, one must assess if the controller or processor has access to data points that alone or in combination could identify the data subject.[53] In the negative, one must objectively assess if the controller or processor,[54] or "another [person]" of relevance,[55] has at their disposal a means[56] "reasonably likely to be used"[57] of revealing an identifier.[58]

Reasonable likelihood excludes methods "practically impossible"[59] to employ, that are prohibitively time-consuming, expensive, personnel-draining, or resource-intensive. Unlawful means of re-identification are also excluded.[60] The GDPR and the *Breyer* test are silent to whether means of re-identification with stochastic (randomly fluctuating) probabilities of success are considered. "Reasonable likelihood" refers to the likelihood of a means being deployed – not its probability of success.[61]

We interpret *Breyer*'s consideration of third parties to include a matrix of third parties sufficiently proximate to the controller in relationship, or having a

---

52 *Ibid*, para. 25.

53 Miranda Mourby and others, "Are 'Pseudonymised' Data Always Personal Data? Implications of the GDPR for Administrative Data Research in the UK" (2018) 34 *Computer Law and Security Review* 222-233.

54 Frederik Zuiderveen Borgesius, "The Breyer Case of the Court of Justice of the European Union: IP Addresses and the Personal Data Definition" (2017) 3 *European Data Protection Law Review* 130-137.

55 Lorenzo Dalla Corte, "Scoping Personal Data: Towards a Nuanced Interpretation of the Material Scope of EU Data Protection Law" (2019) 10 *European Journal of Law and Technology*, s. 2.2.

56 *Breyer*, *supra* n. 51, paras. 47-48.

57 GDPR, *supra* n. 41, Recital 26.

58 *Breyer*, *supra* n. 51, para. 32; *Scarlet Extended SA v. Société Beige des Auteurs, Compositeurs et Éditeurs SCRL* Case C-70/10 [2011], para 51; Article 29 Data Protection Working Party, "Opinion 4/2007 on the concept of personal data", *supra* n. 43, pp. 12-15.

59 *Breyer*, *supra* n. 51, para. 46.

60 *Breyer*, *supra* n. 51, paras. 45-49.

61 GDPR, *supra* n. 41, Recital 26.

plausible motive to perform re-identification.[62] We now inquire – what is the underlying "subject" of identification?

2.4 What is an identifier?

Data must be associated to an "identifier" to be personal. We consider what bears a sufficient connection to a natural person to constitute an identifier.

2.4.1 Identifiers in Canadian law

Canadian jurisprudence defines identifiers more conservatively than does EU guidance. Objects (e.g. firearms[63] or vehicles[64]), locations,[65] land descriptions,[66] or transactions,[67] even those bearing a strong link to a natural person, generally cannot be "identifiers", nor can codes associated thereto (e.g. firearm serial numbers).[68] Conversely, drivers' license numbers and vehicle identification numbers are "codes" sufficiently proximate to natural persons to be "identifiers".[69] "Identification" presupposes individuation of the natural person "identified", rather than individuation by way of a "proxy" such as a personal characteristic or identification number. A "proxy" is sufficient only insofar as the proxy inherently permits individuation[70] (i.e., mere knowledge of the proxy equates individuation absent significant effort spent matching the "proxy" to the concerned individual).[71]

---

62 *Nowak*, *supra* n. 46, paras. 30-31.
63 *Safety*, *supra* n. 6, paras. 1-8, 43-48.
64 *Schindler*, *supra* n. 15, paras. 108-112; *Leon*, *supra* n. 14, para. 49; *Otis*, *supra* n. 14, para. 89.
65 *Solicitor*, *supra* n. 14, paras. 19-25.
66 *Alberta-Health*, *supra* n. 21, paras. 58-59.
67 *Attorney General*, *supra* n. 20, paras. 34-43.
68 *Safety*, *supra* n. 6, paras. 43-48.
69 *Leon*, *supra* n. 14, paras. 48-51.
70 *Carleton University, Re* 2013 CarswellOnt 19131, paras. 39-48; *Health*, *supra* n. 15, paras. 17-25; *Halton District School Board, Re* 2001 CarswellOnt 10977, paras. 87-89.
71 *Leon*, *supra* n. 14, paras. 47-51.

For greater clarity, individuation is the process of ascribing data to a specific natural person, even if the nominative identity of the person is unknown. For instance, the process of tracking a singular, but unknown, stranger's movement would constitute individuation. Both Canadian law and the GDPR hold the process of individuation to constitute identification, but the GDPR also explicitly recognises certain identifiers that are wholly external to the concerned natural person.[72]

2.4.2 Identifiers in EU Law

The GDPR "identifier" does not require direct individuation. The identification of a "proxy" uniquely bearing on a singular individual[73] is sufficient, even if the proxy cannot easily be linked to the "identified" natural person.

The broader GDPR identifier can refer to a concept or an object that relates to one individual through the subjective appreciation of a close connection between the individual and the concept or object. For instance, an opinion, genetic code, or household item may be considered the person's identifier because of the subjectively appreciated close connection between the concept or item and the person. This latter form of identifier, the object or concept bearing a subjectively assessed connection to a singular individual, is considered an identifier according to the GDPR but not according to Canadian law.

"Online identifiers" of the electronic devices belonging to natural persons, such as cookies or IP addresses, qualify as GDPR identifiers.[74] Further, data can be both personal information and underlying identifier (i.e., signifier and

---

[72] GDPR, *supra* n. 41, preamble at para. 30.

[73] *Breyer*, *supra* n. 51; *Nowak*, *supra*, n. 46, paras. 33-35; Michèle Finck and Frank Pallas, "They Who Must Not Be Identified—Distinguishing Personal from Non-Personal Data under the GDPR" (2020) 0 *International Data Privacy Law* 11-36.

[74] GDPR, *supra* n. 41, preamble at para. 30.

signified), as rich data conveys personal facts, and also uniquely concerns the underlying individual. Genetic and biometric data are quintessential examples. The GDPR "identifier" can relate to "physical, physiological, genetic, mental, economic, cultural or social identity"[75] absent personal individuation, whereas, in Canada identification is tantamount to individuation.

Consequently, some data anonymised in Canada may be inherently impossible to anonymise in the EU. This because such data (e.g., DNA, biometrics) are the underlying EU identifiers subject to the identification risk, not simply the data objects used to effect re-identification.[76] Further, the broader range of "identifiers" recognised in the GDPR suggests that data held in the EU is more readily deemed 'personal data' than data held in Canada.

2.5 Practical difficulties in the application of Canadian and European identifiability standards – risks for cross-border health data sharing

Identifiability in Canada and the EU each present a risk-based model, predicated on binary identifiability and contextually assessed anonymity.[77] Nonetheless, divergence remains. In performing Canada-EU data sharing, data anonymous in the EU might not always be anonymous in Canada, and vice-versa. The distinctions are as follows.

Canada's "personal" criterion is narrower than its EU counterpart, requiring at maximum a privacy interest in the content of the data, and at minimum that the content of the data relate to the concerned individual. The

---

75 GDPR, *supra* n. 41, art. 4(1).

76 GDPR, *supra* n. 41, preamble at paras. 34-35.

77 Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques" 0829/14/EN WP216 (2014), pp. 8-10.

GDPR's more expansive "personal" criterion requires only a "relation" of the data to the individual as regards content, purpose, or effect.

Canada's 'identifiers' are more limited than their GDPR equivalents. The "identifier" must enable the personal individuation of the concerned natural person, directly or via a proxy equating individuation. Codes, objects, or vehicles generally cannot be identifiers. In the EU, the identifier need not individuate or identify the concerned individual by name; biological features, characteristics, items, and opinions can be identifiers so long as they relate exclusively to one individual.

Different tests assess the reasonably foreseeable prospect of identification. The Canadian test accounts for illicit and inadvertent means of re-identification, and those with stochastic probabilities of success. The GDPR interprets re-identification more expansively, excluding only "practically impossible" means of re-identification. Nonetheless, the GDPR appears to exclude illicit or inadvertent re-identification risks, and its treatment of identification risks of indeterminate potency is yet unknown. The GDPR test is generally broader, but can exclude risks accounted for in Canadian guidance.

The language of statutory and jurisprudential tests for data identifiability in Canada and the European Union is qualitative, as it requires regulated parties, regulators, and judges to ascertain the personal nature of an identifier, the personal nature of data and/or the relation of the data to an identifier, and the risk of re-identification according to jurisprudential criteria that are not expressed using empirically measurable standards. This creates two risks.

The first risk is that the qualitative articulation of foundational concepts in data privacy law can make compliance difficult even for well-intentioned regulated parties. If the personal and identifiable nature of data is subjectively appreciated, it cannot be meaningfully related to specific engineering practices

to be implemented by the designers of informatics systems and by data managers responsible for the curation of databases.

Rubinstein[78] and others[79] acknowledge a similar difficulty in their comparison of legal conceptions of "data protection by design" or "privacy by design" to the technological methods used to actualise such legal standards. These authors conclude that the determination that data protection or privacy "by design" has been adhered to has generally constituted an exercise in ex-post rationalization by regulators. That is, no appreciable relationship exists between the legal concept of data protection or privacy "by design" and the practices regulators have approved as compliant with principles of data protection by design, or privacy by design.[80]

The second – and closely related – risk is that ambiguity as to the data that is captured by data privacy legislation can decrease the overall certainty of the regulatory regime. Uncertainty arises because of the inherent indeterminacy of the qualitative language used to articulate the legal boundaries of personal data. This is especially challenging for potentially regulated entities, as data identifiability is determinative of the application of the entire regime of data privacy or data protection law. If entities err in their assessment of data identifiability, it is possible for them to unintentionally engage in considerable non-compliance.

This challenge is magnified in the context of interjurisdictional data exchange. Courts in Canada and the European Union have expressed generally

---

78 Ira S. Rubinstein and Nathan Good, "Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents" (2013) 8 *Berkeley Technology Law Journal* 1333. Ira S. Rubinstein, "Regulating Privacy by Design" (2012) *Berkeley Technology Law Journal* 26, 1409.

79 Avner Levin, "Privacy by Design by Regulation: The Case Study of Ontario" (2018) 4 *Canadian Journal of Comparative and Contemporary Law* 115.

80 Rubinstein and Good, *supra* n. 78; Levin, *supra* n. 79.

similar conceptions of data identifiability, but have applied such conceptions divergently and reached different conclusions in jurisprudence as to the categories of data that can be considered personal. Such outcomes reinforce our thesis that qualitative expressions of data identifiability are inherently ambiguous, and open to changing interpretations – to the detriment of entities attempting best-efforts compliance. This misalignment can create inefficiencies in the international exchange of data, especially as it can be difficult to determine which of many potentially overlapping data privacy or data protection statutes are applicable to a particular data processing activity.

Further, entities could potentially be subject to competing duties to disclose data considered anonymised according to the laws of one jurisdiction (i.e., according to access-to-information legislation or clinical trial data disclosure requirements) and to withhold the same data on the grounds that it is identifiable personal data according to the laws of another jurisdiction. In the following section, we will demonstrate the limitations of data privacy and data protection laws in addressing a broad range of harms potentially caused by data use, that can arise irrespective of the identifiable or non-identifiable character of the data processed.

Data privacy laws are increasingly attempting to regulate the use of technologies to perform algorithmic profiling, automated decision-making, behavioral modification, and surveillance.[81] Numerous authors have demonstrated that the relationship between delimited categories of regulated "personal" data and the data-related practices that law hopes to regulate is

[81] Article 29 Data Protection Working Party "17/EN WP 251 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" (2017); Office of the Privacy Commissioner of Canada "PIPEDA Case Summary #2002-42: Air Canada allows 1% of Aeroplan membership to 'opt out' of information sharing practices" (2002).

increasingly tenuous.[82] That is, practices that can challenge individual rights to privacy and informational self-determination may no longer require the use of "personal" or "identifiable" data that bears an appreciable relation to a discrete individual.[83]

Such conceptual limitations are thoroughly described in the following section. We contend that these conceptual limitations may further impel regulators and courts to interpret the concept of personal data according to policy imperatives rather than self-consistent legal standards. This may threaten the coherence and predictability of data privacy law.

3 Conceptual challenges for legal "identifiability" in informatics and information philosophy literatures

Canadian law does not capture information without substantive content associable to an identifiable individual. The GDPR can capture information without substantive content associable to an individual by virtue of its "purpose" or "effect" prongs.[84] Its theoretically universal ambit is restrained by the identifiability criterion.[85] The GDPR adopts a "very broad implicit adversarial model"[86] but does not govern data processing that can "generate informational harm"[87] without a specific individual being "identifiable" nor circumstances

---

82 Ignacio Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70 *Hastings Law Journal* 1390, pp. 1412-1424; Dara Hallinan, "Data Protection without Data: Data Protection without Data: Could Data Protection Law Apply without Personal Data Being Processed?" (2019) *European Data Protection Law Review* 293; Dara Hallinan et al., "Neurodata and Neuroprivacy: Data Protection Outdated?" (2014) 12 *Surveillance & Society*, S. 55–72.

83 Alessandro 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* 165-177.

84 *Nowak*, *supra* n. 46, para. 34.

85 Purtova, *supra* n. 17, p. 54; Dalla Corte, *supra* n. 55, s. 4.2.2.

86 Dalla Corte, *supra* n. 55, s. 3.3.

87 *Ibid*, s. 4.2.2.

wherein one can "infer meaningful attributes"[88] about a group absent individual identifiability.

The limitations of legal "identifiability" metrics will be considered with reference to the (a) informatics, and (b) information philosophy literatures. Dalla Corte and Purtova rely on such literatures to demonstrate that a formalist reading of data protection law's material scope suggests that personal data is all-encompassing.[89] Dalla Corte argues that accounting for the GDPR's principles can establish sensible boundaries thereto (the analysis remains cogent in Canada).[90] We contend that data privacy law's identifiability metrics are simultaneously over-inclusive and under-inclusive of their stated regulatory targets. These include data communication, information inference, and the use of autonomous "algorithmic" agents to affect natural persons.[91]

3.1.1 Informatics Literature

Informatics literature discusses three principal forms of identification. The first, "identity disclosure", or "identity inference" is the linkage of an individual identifier to that individual's corresponding record in the dataset. This can refer to a presumptively anonymised health record being linked to the concerned patient, and thus re-identified.[92] Identity disclosure is the principal identification

---

88 *Ibid*.

89 Purtova, *supra* n. 17, at p. 54; Dalla Corte, *supra* n. 55, s. 4.2.2.

90 Dalla Corte, *supra* n. 55, ss. 4.2, 4.3.2.

91 Government of Canada, "Fact Sheet: Digital Charter Implementation Act, 2020" (2020); Office of the Privacy Commissioner of Canada, "Consultation on the OPC's Proposals for Ensuring Appropriate Regulation of Artificial Intelligence" (2020); Innovation, Science, and Economic Development Canada, "Strengthening Privacy for the Digital Age", Government of Canada (2019); Office of the Privacy Commissioner of Canada "Guidance on inappropriate data practices: Interpretation and application of subsection 5(3)" (2018); Office of the Privacy Commissioner of Canada, "Report on the 2010 Office of the Privacy Commissioner of Canada's Consultations on Online Tracking, Profiling and Targeting, and Cloud Computing" (2011).

92 Khaled el Emam et al., "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records" (2009) 62 *Canadian Journal of Hospital Pharmacy* 307-319, p. 309.

risk contemplated in law.

In genomics, this approach enables re-identification attacks on genomic Beacons, wherein limited genomic information from a target individual is compared to the limited genomic information in the genomic Beacon to infer the individual's presence or absence.[93] A genomic Beacon is a system that returns binary answers to queries for limited genetic information, such as the presence or absence of a specified genetic variant across the aggregated data in a bank of genomic records. Membership inference is performed by successively querying the known genomic variants of the target to infer whether the queried bank contains their record. Identity disclosure refers to the association of an individual's identity (i.e., identifier) to a specific record. Membership inference is a closely related concept; it refers to the inference that a specific individual's record is comprised in a dataset without necessarily being able to distinguish their specific record within the larger dataset.

The second, "attribute disclosure", or "attribute inference" confirms that a particular attribute is represented among the records of a database, without necessarily establishing to which records the attribute pertains.[94] Canadian law would not capture such a risk. Attribute disclosure or inference may be captured by the GDPR if the attribute has been associated to an 'identifier'.

The third, "record matching" or "attribute matching", compares "quasi-identifiers" or attributes across datasets. It does not intend the definite

---

93 JL Raisaro et al., "Addressing Beacon Re-Identification Attacks: Quantification and Mitigation of Privacy Risks" (2017) 24 *Journal of the American Medical Informatics Association* 799-805; Suyash Shringarpure and Carlos Bustamante, "Privacy Risks from Genomic Data-Sharing Beacons" (2015) 97 *American Journal of Human Genetics* 631-646.

94 El Emam et al., "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records", *supra* n. 92; Sahel Shariati Samani, "Assessing Disclosure Risks with Genomic Data" (Doctoral Thesis, University of Manchester 2018), pp. 41-50; Athanasios Andreo, Oana Goga and Patrick Loiseau, "Identity vs. Attribute Disclosure Risks for Users with Multiple Social Profiles" (2017) 17 *Proceedings of ASONAM* pp. 2-5.

identification of a record or attribute, but probabilistically assesses the likelihood that the records or attributes composing two distinct datasets overlap.[95] In computer science, such an approach (a "model inversion" or "membership inference" attack)[96] can enable the source code of a machine learning algorithm or neural network[97] to re-identify the input data that "trained" it. Canadian law and the GDPR presumably treat the inference that an individual's data has a probabilistic chance of being comprised in a dataset as personal data. Knowledge that two data pools have a certain probability of containing overlapping data, attributes, or records is not captured in data protection law's scope.[98] "Identifiability" metrics thus fail to account for "informational harms" caused by data that is non-individuating, and by the "relation" of data to things (e.g., data pools, attributes, groups) not considered "identifiers".[99]

3.1.2 Information Philosophy

Information philosophy questions the normative content of "data" and "information", drawing from computer science, information theory, linguistics, mathematics, and philosophy.

Establishing individual "identifiability" and personal "content" (Canada) or "relation" (EU) are evidentiary precursors to the application of data protection

---

95 Gregory Simon et al., "Assessing and Minimizing Re-Identification Risk in Research Data Derived from Health Care Records" (2019) 7 *eGEMS* 1, pp. 4-9; Vinenç Torra, "Privacy Models and Disclosure Risk Measures" (2017) 28 *Data Privacy: Foundations, New Developments, and the Big Data Challenge,* ss. 5.4-5.9.

96 Michael Veale, Reuben Binns and Lilian Edwards, "Algorithms That Remember: Model Inversion Attacks and Data Protection Law" (2018) 376 *Philosophical Transactions of the Royal Society* 1-15, pp. 5-6.

97 Reza Shokri and others, "Membership Inference Attacks Against Machine Learning Models" *IEEE Symposium on Security and Privacy, Proceedings 2017* (2017), pp. 4-6.

98 Veale et al., *supra* n. 96, pp. 3-9.

99 Brent Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate" (2016) 3 *Big Data and Society* 1-21, pp. 6-12.

law. "Semantic" human intelligibility is thus an implicit conceptual boundary of personal data. Such framing can be traced to the emergence of data protection law in the context of information communication technologies (ICTs), which communicate information and organise data, the semantic content of which is comprehensible.[100]

Identifiability metrics are sufficient to regulate the use of data by humans and ICTs, as courts or regulators can assess the "relation" of data to an individual, alone or in collaboration with experts. Data protection law struggles to regulate algorithms. In keeping with Mittelstadt et al., we use "algorithms" to denote those computational mechanisms that organise information, create profiles, and autonomously act in fashions that are unintelligible to humans and often autopoietic (algorithms that can self-modify) and relegate manually-designed and human-comprehensible "algorithms" to the category of communicative ICTs.[101]

Mathematical communication theory (MCT) and "levels of abstraction" (LoA) can concretise the conceptual difficulties of data-centric algorithm regulation.

MCT quantifies the communicative potential of data to measure the number of potential interpretations of a data communication. MCT divides a communicated "message" into composite "strings" composed of "bits", each "bit" reflecting a certain "symbol" among the range of "symbols" that can be expressed.[102] The informational potential of a "bit" is defined by the potential "symbols" it can express. E.g. a letter in the English alphabet can express 26

---

[100] Raphaël Gellert, "Data Protection and Notions of Information: A Conceptual Exploration" (2019) Working Paper, pp. 3-10.

[101] Mittelstadt et al., *supra* n. 99, p. 3.

[102] Luciano Floridi, "Philosophical Conceptions of Information" (2009) 5363 *Formal Theories of Information Lecture Notes in Computer Science*, 13-53, pp. 25-35.

possible "symbols".[103] The "informational" potential of a string can be calculated as a function of the number of possible combinations of "bits" in a given "string".[104] MCT is useful for assessing the *number* of potential different meanings a message of a determinate length using a determinate code *could* take on, and for calculating the probability of the message being *corrupted*.[105] MCT does not interest itself in the semantic content of a message at all. The successful transmission of a *semantic* message requires the transmitter and interpreter's agreement as to the context of interpretation.[106] The implication, ironically, is that while the expressive potential of data can be quantified within an agreed symbolic-representational system, the semantic meaning of communicated data arises independently from its constituent data and cannot be inferred therefrom.

What does this imply for data protection? For the regulation of communication using traditional ICTs, the "relating to" criterion is appropriate as humans can determine if data semantically "relates to" persons with a measure of success. In regulating algorithms, the input data and output data as apparent to humans are immaterial, as the inferences the algorithm draws from data bear no semantic relationship to the inferences humans draw therefrom.[107]

---

[103] *Ibid*, pp. 26-28.

[104] *Ibid*, pp. 31-32; Olivier Rioul, "This Is IT: A Primer on Shannon's Entropy and Information" (2018) *XXIII L'Information, Séminaire Poincaré*, pp. 49-61; Nimrod Bar-Am, "Revisiting Context and Meaning: Claude Shannon's Mathematical Theory of Communication" in Bar-Am (ed.) *In Search of a Simple Introduction to Communication* (Springer: 2016), pp. 123-140.

[105] Floridi, "Philosophical Conceptions of Information", *supra* n. 102, p. 31. Loet Leydesdorff, Mark Johnson and Inga Ivanova, "Toward a Calculus of Redundancy: Signification, Codification, and Anticipation in Cultural Evolution" (2018) 69 *Journal of the Association for Information Science and Technology* 1181-1192.

[106] Sabina Leonelli, "The Philosophy of Data" in Luciano Floridi (ed.), *The Routledge Handbook for the Philosophy of Information* (Routledge: 2016), pp. 19-21.

[107] Hermann Kopetz, "Information Versus Data," in Kopetz (eds.) *Simplicity is Complex: Foundations of Cyber-Physical System Design* (Springer International: 2019) 19-36, pp. 20-25; Martin Thellefsen, Torkild Thellefsen, and Bent Sørensen, "Information as Signs: A Semiotic Analysis of the Information Concept, Determining Its Ontological and Epistemological Foundations" (2018) 74 *Journal of Documentation* 372-382.

Further, algorithms often model complex systems of phenomena, rather than communicating and replicating phenomena as do ICTs; consequently the workings of an algorithm are often revealed in considering the interrelationships between "data" rather than the unitary characteristics thereof.[108]

"Levels of abstraction" (LoA) can represent systems by selecting "observable" features and using those features to "model" the system.[109] Complex systems can be modeled using levels of "abstraction" that represent differing perspectives on, or levels of granularity of, a selfsame system, with or without common "observables" being used to model each "level".[110]

The same "input data" can be translated into differing models based on the "observables" selected and the "system" modeled. The personal character of data in data protection law refers to "observables" apparent to humans. Conversely, a model might be described as "personal" data if the "observables" it defines closely relate to human features or behaviour, or if the "system" modeled approximates human activities. These elements are agnostic to the personal or impersonal nature of input data.[111] For data protection law, defining data as "personal" or "impersonal" from a human-intelligible perspective provides no insight into the personal or impersonal character of the "observables" an algorithmic model relies on and the system it models.[112]

---

[108] Douglas Marsh, "Toward a Phenomenology of Information: Philosophical Engagements with Information Technology in the Information Age" (2018) 8 *Prince Songkla University Journal of International Studies*, s. 3.2.1.

[109] Floridi, "Philosophical Conceptions of Information", *supra* n. 102, pp. 36-38.

[110] Luciano Floridi, "The Method of Levels of Abstraction" (2008) 18 *Minds and Machines* 303-329, pp. 309-316.

[111] Luciano Floridi, "Group Privacy: A Defence and an Interpretation" in Linnet Taylor et al. (eds.), *Group Privacy: New Challenges of Data Technologies* (Springer International: 2017) 83-100, pp. 85-90.

[112] Luciano Floridi, "The Logic of Design as a Conceptual Logic of Information" (2017) 27 *Minds and Machines* 495-519, pp. 497-503.

A system modeled might consist of "group behaviour" or "material behaviour" inferred from "observables" of the collective or the material world that do not relate to individuals.[113]

Mittelstadtian algorithms are unconcerned with data as a representation of the world, but rather, are meaning-generative, creating from data knowledge not enshrined in its content.[114] ICTs attempt the representation of data's ontic elements and the symbolic translation thereof. Algorithmic technology attempts the induction of structural ontological elements from information not contained in the data, but inferred from their interrelations.[115]

The implications for data protection and data privacy law are as follows. The General Data Protection Regulation has integrated many novel provisions that are applicable to automated decision-making.[116] The territorial scope of the GDPR,[117] the Recitals that accompany the GDPR, and regulatory guidance applicable thereto[118] all reiterate that behavioral monitoring and profiling are central regulatory targets of data protection law. The law's material scope, conversely, limits its application to "identifiable personal data".

EU law attempts to regulate the aforementioned practices by protecting individuals against automatic decision-making (subject to exception),[119] and

---

113 Floridi, "Group Privacy: A Defence and an Interpretation", *supra* n. 111. Michele Loi and Markus Christen "Two Concepts of Group Privacy" (2019) 1 (forthcoming) *Philosophy and Technology.*

114 Gellert, *supra* n. 100, pp. 11-19.

115 Erik Radio and others, "Manifestations of Metadata Structures in Research Datasets and Their Ontic Implications" (2018) 17 *Journal of Library Metadata* 161, pp. 176-178; Serge Gutwirth and Mireille Hildebrandt, "Some Caveats on Profiling" in Serge Gutwirth et al. (eds.), *Data Protection in a Profiled World* (Springer: 2010), pp. 32.

116 GDPR, *supra* n. 41, art. 22.

117 GDPR, *supra* n. 41, art. 3.

118 GDPR, *supra* n. 41, Recital 71; Article 29 Data Protection Working Party "17/EN WP 251 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679", *supra* n. 81.

119 GDPR, *supra* n. 41, arts. 4(4), 22.

granting individuals limited informational rights in automatic decisions affecting them.[120] Canada grants individuals more robust informational rights, applicable to the automated administrative decisions of the State. These include access to the source code of decisional algorithms, guarantees of data quality, obligations to screen for biased outcomes, and an explicit right to decisional explanation broader than that recognised in EU law.[121]

The GDPR applies only to the processing of personal data. Conversely, its regulatory targets include practices that are agnostic to the processing of personal data, as demonstrated by our foregoing analysis.[122] Data protection and data privacy laws attempt to regulate practices that include algorithmic decision-making and individual profiling practices. However, such practices may often occur in circumstances where relevant law may not be applicable (i.e., no demonstrably personal data is processed); or where the law should find application but this is difficult or impossible for human observers to establish (i.e., an algorithm processes personal data, but this is not possible for human auditors to ascertain on reviewing the algorithm's behavior).

The qualitative framing of the law's data identifiability standards thus leaves ambiguous whether regulators will interpret "identifiable personal data" broadly to assert jurisdiction over data uses that could be harmful to individuals or to society – or alternatively, whether data protection and data privacy law will

---

120 *Ibid*, arts. 12(3), 13-15; Margot Kaminski, "The Right to Explanation, Explained" (2019) 34 *Berkeley Technology Law Journal* 189-218.

121 Government of Canada, "Directive on Automated Decision-Making" (2019).

122 Cofone, *supra* n. 82. Brent Mittelstadt, "From Individual to Group Privacy in Biomedical Big Data" in I Cohen et al. (eds.) *Big Data, Health Law, and Bioethics* (CUP, 2018), pp. 176-185; Sandra Wachter, Brent Mittelstadt and Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR" (2018) 31 *Harvard Journal of Law and Technology* 841-887, pp. 859-860; Sharona Hoffman, "Big Data's New Discrimination Threats: Amending the Americans with Disabilities Act to Cover Discrimination Based on Data-Driven Predictions of the Future" in I Cohen et al. (eds.), *Big Data, Health Law, and Bioethics* (CUP, 2018), p. 90.

simply not apply to data uses that adversely impact individuals and society, as no identifiable personal data has demonstrably been processed.[123]

For greater clarity, the difficulty is as follows. In processing data, regulated entities must assess which data used is identifiable personal data and which data is not. This assessment must be performed as data privacy and data protection law impose considerable procedural and substantive obligations on the use of identifiable personal data, and the failure to treat identifiable personal data as such can expose regulated entities to considerable risk of non-compliance (i.e., with obligations to obtain consent or define a lawful basis for data use, to maintain necessary records, to respect ongoing individual rights, etc.). Therefore, it is beneficial for regulated entities to be able to easily grasp what data is identifiable personal data, as the correct assessment thereof is a precondition to compliance with the law's substantive and procedural requirements which are applicable to all identifiable personal data.

For regulated entities to better be able to assess and ensure their compliance with the law, only demonstrably identifiable data should be considered personal data. Conversely, many potentially objectionable uses of data that affect individuals can be performed without the use of any data that demonstrably relates to those specific persons from the perspective of the controller or processor (e.g., dynamic IP addresses in *Breyer*). As a result, courts have an incentive to interpret "identifiable data" liberally so as to provide

---

[123] El Khoury, *supra* n. 80, describes the broad definition of personal data for the purposes of asserting jurisdiction on a processing activity. El Khoury contends that, in *Breyer*, the court asserted that data was identifiable on the basis of powers possessed by, and data available to, third parties other than the concerned data controller. The difficulty in inherent in such a position is that the legal powers of third parties, and the data available to such parties for the purposes of effecting re-identification cannot be known prospectively by data controllers that process data for their own purposes. Consequently, it would appear that data controllers may not always have the necessary information to know if the data they process is personal data or not (and therefore, whether or not they are subject to data protection legislation).

remedies concerning potentially objectionable practices that use data that is not demonstrably identifiable. However, the result of such a broadening is that data controllers and data processors can find themselves in breach of obligations relating to data that they could not possibly have known to be identifiable personal data.

The tension observed here is that it is not possible to narrow the interpretation of identifiable data without precluding courts from intervening to regulate objectionable uses of non-identifiable data. Conversely, accepting a broad or shifting definition of identifiable data risks exposing entities that use data to considerable regulatory risk and procedural burdens applicable to all of their uses of data.

Therefore, adopting a narrower definition of personal data is justified if other supplementary legal regimes can regulate potential misuses of data unrelated to individual identification. For the health sector, this would better enable the use of rich health data that is not identifiable for the purposes of quality assurance activities and quality assurance activities. Access to rich health data is instrumental to the further implementation of precision medicine and computational approaches to healthcare delivery. The present approach prompts conservatism in secondary data use, limiting the potential for the health sector to translate the data at its disposal into research outputs, clinical decision tools, and accurate predictions regarding healthcare outcomes. A more consistent and more liberal definition of anonymised data would enable the health sector to benefit from increased mobility of data and secondary data use, allowing for improved research and better public-health decision-making without compromising individual privacy.

Our first proposed solution to these challenges is to conceptually distinguish the legal recognition of individual interests in data privacy and data

protection from the regulation of data uses more generally, and of profiling and surveillance practices.

This first solution is gaining traction amongst Canadian and EU regulators. In Canada, proposed amendments to the private-sector data privacy law extend some protections to de-identified personal data.[124] The *Directive on Automated Decision-Making* applies to automated decision-making and is agnostic to the personal character of any data implicated.[125] Neither the proposed EU framework to regulate artificial intelligence ethics,[126] nor the draft *Data Governance Act*[127] is constrained in scope to identifiable personal data.

Our second proposed solution is for regulators and regulated parties to express the identifiability of their data, and other ethico-legal characteristics of their data (e.g., data utility, potential for data processing to reveal group-specific attributes) relative to measurable and formally expressed standards. The use of such standards can incentivise best-efforts compliance by creating clear metrics against which to measure compliance efforts. In the following section, we describe emergent quantitative methodologies used to express and measure the risk of individual re-identification and other privacy-related risks. We contend that these methodologies could be incorporated to future legislation, regulatory guidance, or sectoral codes of conduct to help guide entities in assessing privacy risks and in producing evidence of their data privacy compliance efforts. Prior to

---

124 Bill C-11. An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts. *Digital Charter Implementation Act*, 2020, at ss. 62(2) and 63(3).

125 Government of Canada, "Directive on Automated Decision-Making", *supra* n. 121.

126 European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL). Framework of ethical aspects of artificial intelligence, robotics and related technologies.

127 European Commission. Proposal for a Regulation of the European Commission and of the Council on European data governance (Data Governance Act) COM/2020/767.

the enshrinement of such metrics in regulatory guidance or legislation, these instruments will nonetheless be useful for data controllers and data users to justify their legal compliance practices relative to objective benchmarks. In the following section, we provide an overview of existing quantitative and empirical approaches to evaluating data identifiability and articulating other concepts related to data governance and associated regulatory compliance.

4 Towards Quantitative Models for Assessing Data Identifiability

4.1 Existing use of quantitative models for assessing the "re-identification risk" in health data

Canadian[128] and European[129] health authorities have issued "de-identification" guidance for clinical trial data applicable prior to its public release. The clinical trial data release policies of health authorities in Canada and the EU and the regulatory guidance of Canadian privacy regulators establish maximum percentile re-identification risk scores for data to be considered anonymous or anonymised. A nine percent maximum re-identification risk threshold is generally proposed for data to be considered "anonymised".[130]

We interpret as follows the relationship between the data identifiability standards enshrined in case law, and the maximum percentile re-identification risk scores described in health agency guidance and data privacy regulatory guidance. The qualitative tests enshrined in case law constitute the legally

---

[128] Health Canada, "Guidance Document on Public Release of Clinical Information" (2019).

[129] European Medicines Agency, "External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use" (2018), pp. 39-43, 49.

[130] *Ibid*. Health Canada, *supra*, n. 128.

binding tests that courts use to rule as to whether data is identifiable personal data or not. The maximum percentile risk scores proposed by regulators are interpretations of the jurisprudential tests, which do not supersede the jurisprudential tests in law's hierarchy of norms.

Regulators further propose methodologies for performing data de-identification and calculating residual data re-identification risks. These are proposed methods for rendering data anonymised, but do not supplant the established jurisprudential tests for assessing data identifiability.

"K-anonymisation" is often used in health sector regulatory guidance to measure structured data's percentile re-identification risk. "K-anonymisation" considers how many records within a record-set exhibit the same combination of quasi-identifiers. Quasi-identifiers are attributes that alone do not identify individuals, but could through their combination.

4.1.1 Policy justifications for re-identification risk quantification

If a dataset has three "dimensions" of quasi-identifiers (e.g., age, gender, and profession), the percentile re-identification risk is expressed as = "1/k", where "k" is the smallest "equivalence class". An equivalence class is the number of individuals in a dataset sharing the same combination of quasi-identifiers. In our above example, if at least eleven individuals comprised in a dataset shared each combination of age, gender, and profession represented, the calculation performed would be 1/11 = percentile risk. Eleven is the smallest equivalence class, and thus our "k" value. 1/11 = 0.09. The re-identification risk is nine percent.

Suppressing records and generalising variables can increase the "k" value. If a dataset has a low "k" value because each age is only represented a few times, the "k" value can be increased, through generalisation, by representing the ages

as intervals such as 1-5, 5-10, and 10-15. Suppression can be achieved by eliminating records comporting outlier ages.[131]

The European Medicines Agency and Health Canada policies consider datasets anonymised if at least eleven records compose each "equivalence class", These policies recommend anonymising data using de-identification processes that minimise identifiability and maximise utility[132] Redaction – deletion of data-fields – is discouraged as it mars data utility.[133] Similar methodologies for assessing data identifiability have been proposed in the regulatory guidance of the Office of the Privacy Commissioner of Canada and the Office of the Privacy Commissioner of Ontario.[134]

Critics of "k-anonymisation" highlight that it cannot anonymise highly dimensional datasets or datasets comprising uncommon quasi-identifiers.[135] A highly "dimensional" dataset is one that exhibits many attribute fields for each record. It is difficult to "k-anonymise" a highly dimensional dataset because the many attributes likely include many quasi-identifiers, thus the "equivalence classes" of individuals exhibiting the same combination of quasi-identifiers will be small.[136]

---

131 Khaled El Emam and Fida Dankar, "Protecting Privacy Using K-Anonymity" (2008) 5 *Journal of the American Medical Informatics Association* 627-637.

132 Jean-Marc Ferran and Sarah Nevitt, "European Medicines Agency Policy 0070: An Exploratory Review of Data Utility in Clinical Study Reports for Academic Research" (2019) 19 *BMC Medical Research Methodology*.

133 Timo Minssen, Neethu Rajam and Marcel Bogers, "Clinical Trial Data Transparency and GDPR Compliance: Implications for Data Sharing and Open Innovation" (2020) 47 *Science and Public Policy*, 616-626, pp. 620-623.

134 Health Canada, *supra* n. 128. European Medicines Agency, *supra* n. 129; Information and Privacy Commissioner of Ontario, *supra* n. 29; Khaled El Emam and others, "Pan-Canadian de-Identification Guidelines for Personal Health Information" *supra* n. 30.

135 Krista Wilkinson and others, "Less than Five Is Less than Ideal: Replacing the "Less than 5 Cell Size" Rule with a Risk-Based Data Disclosure Protocol in a Public Health Setting" (2020) 111, *Canadian Journal of Public Health*, 761-765.

136 Naoise Holohan et al., "(k, e)-Anonymity: k-Anonymity with e-Differential Privacy," (2017) *IBM Research – Ireland*, p. 4.

Reliance on k-anonymisation can disproportionately exclude rare disease cohorts, indigenous groups and minority ethnic groups, and individuals with intersectional identities from health research participation. Their data will often be excised from datasets because their quasi-identifiers are more unique than those of individuals from the dominant culture.[137]

4.2 Re-identification risk quantification

4.2.1 Policy justifications for re-identification risk quantification

Expressing data identifiability standards as maximum re-identification risk thresholds, calculated using specified methodologies will ensure the predictability of the law. This innovation would facilitate compliance with competing legal obligations to disclose data and to withhold data from disclosure. The law presently requires institutions to release data in certain instances. For instance, public institutions have legal obligations to release information to the public on request, after anonymising the data and severing all remaining personal information.[138] Clinical trial regulations also require the public release of clinical trial data in anonymised format. However, it would contravene the law for regulated entities to release personal information in complying with either of these disclosure obligations.[139]

Consequently, the qualitative character of data identifiability metrics leaves regulated entities in a precarious position. Failure to disclose data that is not identifiable personal data could constitute legal non-compliance; disclosing

---

[137] Wilkinson, *supra* n. 135.

[138] *Attaran v. Canada (Minister of National Defence)*, [2011] F.C.J. No. 836, 2011 FC 664.

[139] Matthew Mayernick, "Open Data: Accountability and Transparency" (2017) 4 *Big Data and Society* 1-5; Meg Young and others, "Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing", *Conference on Fairness, Accountability, and Transparency* (2019) 191-200; Council of the Canadian Academies, *supra* n. 5, pp. 152-153.

data that is identifiable personal data could also constitute legal non-compliance. The boundary between these categories is amorphous; therefore, deciding which data to withhold and which data to release can appear an arbitrary determination. Using stipulated maximum re-identification risk scores in combination with stated data de-identification methodologies and re-identification risk assessments could increase accountability and simultaneously facilitate compliance.

It is consequently recommended that legislators and regulators adopt specified maximum re-identification risk thresholds that distinguish identifiable personal data from non-identifiable unregulated data. The maximum re-identification risk thresholds should be supported by approved methodologies for the de-identification of data, to transform identifiable personal data into anonymised unregulated data. Such risk thresholds could further serve to harmonise data privacy compliance practices in the health sector. Presently, a lack of consensus as to the acceptable maximum re-identification risk thresholds in the health sector has caused healthcare entities to tend toward conservatism in their data releases and to adopt zero-tolerances policies regarding residual data re-identification risk.

De-identification methodologies should be structured as soft law instruments, such as guidelines or recommended approaches. The use of soft law to propose data de-identification methodologies allows for regulated parties to propose or to use alternate data de-identification methodologies that are better tailored to the particular features of their datasets and the re-identification risks inherent thereto.

Regulated entities and organisations representing the interests thereof should also proceed to propose such methodologies and obtain regulatory

approval thereof.[140] In the proceeding section, we will consider methodologies that have been devised to de-identify health sector data and that could be incorporated into future legislative or regulatory standards, or instruments of sectoral self-regulation (e.g., codes of conduct, policies, and guidelines).

4.2.2 Methodologies for re-identification risk quantification

Statistical and algorithmic methods, including the aforementioned "k-anonymity" metric, have been devised to de-identify data and calculate the residual re-identification risk of a presumptively anonymised dataset in scientific literature and the data governance practices of health consortia. Collectively, these methods are referred to as statistical disclosure controls, and often serve to calculate the quantity of personal or sensitive data retained in a dataset or record after data de-identification has been performed. These methodologies can generally be implemented to de-identify structured data (i.e., data that is held in labelled and organised tables). It is necessary for the data to be structured for such methodologies to function best, as the parties performing data de-identification and residual re-identification risk calculation must be able to achieve consensus as to the quasi-identifiers and sensitive data fields within the datasets.

"L-diversity" assesses if sensitive attributes are equitably distributed across all equivalence classes of a dataset. It requires each sensitive attribute to be well-represented in each equivalence class composing the overall dataset. This

---

[140] In Europe, this can be actualised through the regulatory approval of Codes of Conduct that can clarify or specify the application of the GDPR to particular economic sectors or processing activities. In Canada, this can be actualised through collaboration with Privacy Commissioners using mechanisms such as "privacy impact assessments". In the future, this could be actualised through the adoption of sector-specific codes of practice and certification programs if Canadian legislation should come to recognise such innovations, as proposed in the draft "Digital Charter Implementation Act".

protects against attribute disclosure in preventing the association of sensitive attributes to particular combinations of quasi-identifiers.[141] As a reminder, an equivalence class is a group of all records sharing the same potentially identifying attributes (quasi-identifiers). Consequently, the l-diversity metric attempts to ensure that specific sensitive attributes cannot be associated to sub-groups of persons sharing common potentially identifying characteristics within the larger dataset. In practice, this is achieved by ensuring that each equivalence class (i.e., each combination of quasi-identifiers) has at least "*l*" different sensitive attributes represented.[142]

"T-closeness" is an analogue of "l-diversity" that safeguards against attribute disclosure by ensuring that the distribution of sensitive attributes in each "equivalence class" is within a specified "distance" of the distribution thereof in the overall dataset[143] (that "equivalence classes" do not overrepresent sensitive attributes).[144]

---

[141] Keerthana Rajendran, Manoj Jayabalan, and Muhammad Rana, "A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data" (2017) 17 *International Journal of Computer Science and Network Security* 172-177; Michael Kern "Anonymity: A Formalisation of Privacy – l-Diversity" (2013) *Network Architectures and Service*; Ashwin Machanavajjhala et al., "L-diversity: Privacy beyond k-anonymity" (2007) 1 *ACM Trans. Knowl. Discov. Data* 1-52.

[142] For instance, if there are twelve different equivalence classes in a dataset, each equivalence class would need to have at least "l" different sensitive attributes represented. E.g.: If a database of diagnosis information is composed of quasi-identifiers age, sex, ethnicity, and profession, to be "l-diverse" to a degree of three, each represented combination of quasi-identifiers would need to have three different diagnoses represented. This precludes a reader from inferring an individual's sensitive attribute (i.e., their specific diagnosis) from their known quasi-identifiers.

[143] Hongyu Liang and Hao Yuan, "On the Complexity of T-Closeness Anonymisation and Related Problems" (2013) *DASFAA 2013: Database Systems for Advanced Applications* 331-345, pp. 337-339.

[144] This can be useful for datasets where certain potentially sensitive attributes are highly common in the datasets and other sensitive attributes are rare. Consider for instance, a table of the population separating individuals with a rare disease from healthy persons. Many equivalence classes might only have one value (i.e., all individuals are healthy). This would not satisfy l-diversity. However, the table could still be "t-close" if rare disease patients were equitably distributed among the different equivalence classes, relative to their distribution in the overall dataset.

"M-invariance" is an anonymisation technique tailored to frequently updated databases, preventing attackers from inferring the sensitive attributes of records added to or removed from the database by making similar queries over time and noting changes in the results as the database is updated.[145]

Population uniqueness can be assessed, using generative models (algorithms) that determine the likelihood that an anonymised record has been successfully re-identified by determining if its combination of quasi-identifiers is unique in a population. Such models can also be used to calculate the proportion of the population that exhibits a unique combination of quasi-identifiers.[146]

The probability of re-identification being *attempted* has been modelled as a metric of re-identification attempt cost relative to value of re-identification, compared to different degrees of data generalisation.[147] The success thereof is contingent on the accurate estimation of attempt cost and re-identified record value.

The re-identification risk of a data element used in a specified context can be modelled and compared to a maximum re-identification risk score to determine if the data used should be considered identifiable or anonymised. Such a modelling exercise is contingent on a cogent assessment of the risks inherent in the data governance context concerned. The purpose of such an exercise is not to perfectly calculate residual data identifiability, but to compare data identifiability relative to a particular risk threshold, provided that the assumptions made about

---

[145] Xiaokui Xiao and Yufei Tao, "M-Invariance: Toward Privacy Preserving Re-Publication of Dynamic Datasets" (2007) *ACM SIGMOD International Conference on Management of Data,* pp. 689-700.

[146] Luc Rocher, Julien Hendrick, and Yves-Alexandre de Montjove, "Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models" (2019) 10 *Nature Communications* 1-9, p. 3.

[147] Zhiyu Wan et al., "A Game Theoretic Framework for Analyzing Re-Identification Risk" (2015) 10 *PloS One*.

the nature of the data, the nature of its quasi-identifiers, and the circumstances of its use prove correct.

4.2.3 Data de-identification algorithms and data utility metrics

De-identification algorithms can apply multiple anonymisation techniques to data and quantify the resulting data's identifiability and utility.[148] Some methodologies supplement statistical metrics using self-modulating algorithmic (e.g. machine learning) techniques to remove quasi-identifiers or sensitive attributes.[149] Other methodologies address the specific features of the target datasets, for instance accounting for the frequent replication of the same direct identifier in clinical records.[150] Verification of successful anonymisation can be performed using statistical metrics[151] or using experimental methodologies to simulate re-identification attacks.[152]

Standalone metrics have also been devised for determining "data utility", that is, how useful data remains after de-identification. Utility is measured as a proportion of data eliminated from the original dataset, known as information loss, or as a metric of how much the de-identified records reflect the original

---

148 Holohan et al., *supra* n. 136; Florian Kohlmayer et al., "A Flexible Approach to Distributed Data Anonymization" (2014) 50 *Journal of Biomedical Informatics* 62-76, pp. 73-75.

149 Pilar López-Ubeda et al., "Anonymization of Clinical Reports in Spanish: A Hybrid Method Based on Machine Learning and Rule" in Miguel Cumbreras et al. (eds.), *Proceedings of the Iberian Languages Evaluation Forum* (2019) 687-695; György Szarvas, Richárd Farkas and Róbert Busa-Feket, "State-of-the-Art Anonymization of Medical Records Using an Iterative Machine Learning Framework" (2007) 14 *Journal of the American Medical Informatics Association* 574-580.

150 Gregory Simon et al., "Assessing and Minimizing Re-Identification Risk in Research Data Derived from Health Care Records" (2019) 7 *eGEMs* 1-9, pp. 5-6; Martin Scaiano et al., "A Unified Framework for Evaluating the Risk of Re-Identification of Text de-Identification Tools" (2016) 63 *Journal of Biomedical Informatics* 174-183.

151 Fida Dankar et al., "Estimating the Re-Identification Risk of Clinical Data Sets" (2012) 12 *BMC Medical Informatics and Decision Making* 1-15, p. 5; Khaled El Emam et al., "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records" *supra* n. 92.

152 Vanessa Ayala-Rivera et al., "COCOA: A Synthetic Data Generator for Testing Anonymization Techniques" (2016) *Privacy in Statistical Databases* 163-177.

records (as opposed to their data being relocated to another record or eliminated), known as data truthfulness.[153] Reliance on utility and truthfulness metrics can be useful in determining which amongst multiple data de-identification methodologies should be used (i.e., which creates output data that is the most faithful to the input data).

4.2.4 Differential privacy

Differential privacy refers to the mathematical demonstration that the process of generating aggregate or summary data from multiple records does not reveal information unique to a singular constituent record in the dataset. Some such methodologies adjust the results of user queries submitted to a dataset rather than altering the data, often by stochastically adding "noise" (minute, random modifications) to the released data to minimise the release of identifiable data.[154] The central characteristic of differential privacy is that it ensures that data from a single record cannot be inferred from the output of a dataset-level analysis (i.e., from aggregate or summary-level analysis of the dataset).

Differential privacy can be achieved by adjusting a database's response to a query such that data released in response thereto is, at most, influenced by the presence of any given record in the dataset by a factor of "ε".[155] Differential privacy prevents a single record's data from skewing the results returned by a factor of "ε" or more. This prevents re-identification attempted by cross-

---

153 Hyukki Lee et al., "Utility-Preserving Anonymization for Health Data Publishing" (2017) 17 *BMC Medical Informatics and Decision Making* 1-12, pp. 3-4.

154 Alexandra Wood et al., "Differential Privacy: A Primer for a Non-Technical Audience" (2018) 21 *Vanderbilt Journal of Entertainment and Technology Law* 209-275.

155 Julian Holzel, "Differential Privacy and the GDPR" (2019) 3 *European Data Protection Law Review* 184-196, pp. 193-196; Wood and others, *supra* n. 154; Andrew Chin and Anne Klinefelter, "Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study" (2012) 90 *North Carolina Law Review* 1417-1456.

referencing an outlier record's known data against the results of strategically formulated queries, to infer the presence or absence of that record from the deviation of results returned by queries including the target and queries excluding the target.

Adopting an approach to privacy compliance that is founded on differential privacy can protect against the re-identification risks created by Mittelstadtian algorithms that are agnostic to the identifiability of their input data. Such protection is successful because it ensures that all input data used to perform an analysis is protected from disclosure and inference.

Consequently, the use of differential privacy to satisfy legislative privacy requirements also provides privacy guarantees that can protect against algorithmic profiling (i.e., that can ensure that algorithms are not making inferences about specific individuals having contributed to a dataset from aggregate data generated from the dataset). Contrary to the aforementioned modelling exercises, differential privacy methodologies can provide mathematical guarantees of data privacy much as cryptography does for data security. It is circumscribed in use, however, to aggregate or summary data regarding multiple records – it cannot, by its nature, be applied to a single record.

5 Conclusion

Certain concessions must be made as to the prospects of quantitative approaches to the expression of residual re-identification risk in data. No universal statistical method exists to articulate overall residual risk in de-identified data. Tension arises between reducing data identifiability and retaining data utility, which may

inhibit a workable trade-off of privacy and utility.[156]

Despite these shortcomings, quantitative approaches to the assessment of data identifiability may provide more certainty to regulated parties than do qualitative approaches currently do. Divergent conceptions of identifiability across health institutions[157] have prompted reactions ranging from the misuse of anonymisation language to refer to identifiable datasets, to approaches admitting zero residual risk in anonymised datasets which effectively bar the secondary use of health data.[158] Data identifiability is the foundational threshold of data privacy and data protection law. Framing such criterion in qualitative, abstract terms threatens the internal consistency of the entire regime. Quantitative metrics can ensure that regulators and regulated parties conceptualise data privacy from shared premises.

For these reasons, we recommend that legislators, regulatory guidance bodies such as the European Data Protection Board and Canadian Privacy Commissioners propose maximum re-identification risk thresholds to guide data privacy law and data protection law compliance. Such an approach has been adopted by the European Medicines Agency and Health Canada to guide public disclosures of clinical trial data, for instance. Select Privacy Commissioners in Canada have also released context-specific data de-identification guidelines that incorporate quantitative assessments. Increased reliance on quantitative metrics will benefit entities hoping to ensure compliance with data privacy laws by

---

156 Mostafa Langarizadeh, Azam Orooji, and Abbas Sheikhtaheri, "Effectiveness of Anonymization Methods in Preserving Patients' Privacy: A Systematic Literature Review" (2018) 248 *Health Informatics Meets eHealth* 80-87.

157 Ira Rubinstein and Woodrow Hartzog, "Anonymization and Risk" (2016) 91 *Washington Law Review* 703-760, pp. 714-717.

158 David Peloquin et al., "Disruptive and Avoidable: GDPR Challenges to Secondary Research Uses of Data" (2020) 28 *European Journal of Human Genetics* 697-705, pp. 698-699; Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" (2009) 57 *UCLA Law Review* 1701-1777.

giving them a demonstrable standard against which to compare their efforts. These metrics can also benefit the public by creating an objective standard against which the practices of potentially non-compliant entities can be assessed.

It is expected that regulators and regulated entities in different circumstances will propose divergent methodologies to calculate the residual re-identification risk of data. The adoption of different methodologies to perform data de-identification is consistent with the contextual approach to data identifiability. As the identifiability of data is related to the context of data use, it is sensible for the most appropriate de-identification methodology to vary from one data processing context to another.[159]

The key benefit of employing maximum re-identification risk thresholds are that these allow for the meaningful comparison of the expected risk of re-identification inherent in data. Using clearly defined methodologies to perform data de-identification and to evaluate residual data identifiability allows for a full understanding of the methods employed to achieve anonymity and to the strengths and potential limitations of the de-identification mechanisms utilised. Differential privacy mechanisms can further be used to offer a strong guarantee that the aggregate or summary level data of multiple persons will not allow for the re-identification of the individuals whose records are comprised in the dataset.

In conclusion, it is first recommended for legislators to distinguish the regulation of data privacy from the regulation of data-related practices that are agnostic to the personal or non-personal character of the data used. Second, it is recommended for legislators and regulators to adopt maximum re-identification

[159] As demonstrated above, for instance, reliance on methodologies such as k-anonymity is suitable to certain data processing contexts (e.g., structured data exhibiting low dimensionality), but different metrics may be more appropriate to modelling re-identification risk in other circumstances.

risk scores to define the boundaries of identifiable and non-identifiable data. Third, it is recommended for regulators and regulated parties to cooperate in developing methodologies for the de-identification of data that are appropriate to specified data processing contexts.

6 APPENDIX A: Maximum Percentile Re-Identification Risk Thresholds

Issuer	Document	Jurisdiction	Maximum Re-Identification Risk Threshold
Health Canada	Public Release of Clinical Information: guidance document (2019)	Canada	9 %
European Medicines Agency	External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use (2018)	European Union	9 %
Office of the Privacy Commissioner of Ontario	De-identification Guidelines for Structured Data (2016)	Province of Ontario (Canada)	Privacy Invasiveness of Release (High): 5% Privacy Invasiveness of Release (Medium): 7.5% Privacy Invasiveness of Release (Low): 10 %
Office of the Privacy Commissioner of Canada	Pan-Canadian De-Identification Guidelines for Personal Health Information (2007)	Canada	1 % (suggested)
Khaled El Emam & Luk Arbuckle (O'Reilly)	Anonymizing Health Data: Case Studies and Methods to Get You Started (2013)	No jurisdiction	5 % – 9 %
Khaled El Emam (CRC Press, Taylor & Francis)	Guide to the De-Identification of Personal Health Information (2013)	No jurisdiction	Individual record re-identification, acceptable maximum probability: 5-10 % (depending on severity of privacy invasion) Individual record re-identification, highest acceptable average probability: 5-10 % (depending on severity of privacy invasion) Re-identification risk, assuming record linkage: 5-

			20% (depending on severity of privacy invasion)

Figure 1: Maximum Percentile Re-Identification Risk Thresholds in Guidance Documents and Informatics Literature

7 APPENDIX B: Glossary of Data Privacy Language

Concept	Description	Topic
Identifier or direct identifier	An attribute such as a name or whole genome sequence that inherently relates to and identifies the concerned natural person.	Ethico-legal data identifiability standards.
Quasi-identifier or indirect identifier	Attributes such as age, gender, or profession that do not inherently the concerned natural person but could lead to individual re-identification through their combination.	Ethico-legal data identifiability standards.
Personal data	Data that poses a reasonably foreseeable prospect of individual re-identification exceeding a stated threshold.	Ethico-legal data identifiability standards.
Anonymous or anonymised data	Data from which all direct identifiers have been removed and the remaining indirect identifiers create no reasonably foreseeable prospect of re-identification.	Ethico-legal data identifiability standards.
Pseudonymous or pseudonymised data	Data from which all direct identifiers have been removed and replaced with a code or pseudonym. A linkage log or other key is retained to enable re-identification.	Ethico-legal data identifiability standards.
De-identification	The process of reducing data's identifiability, for instance by using governance mechanisms or data manipulation techniques.	Ethico-legal data identifiability standards.
HIPAA Safe Harbor De-Identification	The process of the removing the eighteen direct and indirect identifiers listed at HIPAA § 164.514 for the purposes of rendering data anonymised and unregulated according to HIPAA.	Ethico-legal data identifiability standards.
K-anonymisation	For a dataset to be considered k-anonymised, each combination of potentially identifying attributes that appears in the dataset must be identical across at least *k* records. A numerical value must be selected as the *k* value to perform this analysis. Eleven is a commonly selected *k* value used in regulatory guidance and technical literature.	Quantitative data identifiability metric.
Equivalence class	An equivalence class is a group or records within a dataset that have an identical combination of potentially identifying attributes.	Quantitative data identifiability metric.
L-diversity	A metric used to ensure that each sensitive attribute in a dataset is well-represented among all of the equivalence classes of a dataset. This is used to prevent the inference that persons in a dataset with a certain combination of quasi-identifiers tend to manifest a protected sensitive attribute (i.e. to protect against attribute inference or attribute disclosure).	Quantitative data identifiability metric.
T-closeness	A metric, similar to l-diversity, used to ensure that sensitive attributes in a dataset are well-represented among each equivalence class of a dataset.	Quantitative data identifiability metric.

	T-closeness is a variant of l-diversity. If l-diversity considers the prevalence of sensitive attributes in each equivalence class in absolute terms, t-closeness considers the prevalence of sensitive attributes relative to the prevalence thereof in the entire dataset.	
M-invariance	A metric used to ensure that aggregate data preserves privacy even if aggregate data is published from the same source dataset as it changes through time.	Quantitative data identifiability metric.
Differential privacy	A mathematical approach to ensuring data privacy which demonstrates that individual record-level data cannot be inferred from the summary-level data of the dataset, or from queries made concerning an aggregate dataset. This is often achieved by adding a sufficient amount of noise to the results of queries or during the data aggregation process to ensure that data from the record of a single individual will not significantly affect the output of aggregate or summary results from the overall dataset.	Quantitative data identifiability metric.
Dimensionality	Highly dimensional datasets contain more variables or attributes relative to each record. Such datasets are difficult to render anonymous using traditional statistical disclosure controls due to the heterogeneous composition of the data and quasi-identifiers comprised in each record.	Informatics and information philosophy language.
Disclosure	Event directly revealing information about an individual or a dataset.	Informatics and information philosophy language.
Inference	Statistical or algorithmic technique used to probabilistically determine information about an individual or a dataset.	Informatics and information philosophy language.
Identity inference / disclosure	The disclosure or inference that a specific record in a dataset relates to a particular individual.	Informatics and information philosophy language.
Membership inference / disclosure	The disclosure or inference that a specific dataset contains the record of a particular individual.	Informatics and information philosophy language.
Attribute inference / disclosure	The disclosure or inference of an attribute's presence in a dataset or record, or of the prevalence of a certain attribute in the records of a dataset.	Informatics and information philosophy language.
Attribute / record matching	The use of probabilistic inferences and statistical techniques to determine overlap in attributes of multiple records or the attributes comprised in multiple datasets.	Informatics and information philosophy language.

Figure 2: Table of Relevant Terminology

8 APPENDIX C: Comparative Table of Canadian and European Union Data Privacy Law

Concept	Canada	European Union
Identifier	Requires individuation of the concerned natural person. This usually requires identification by name, or with reference to a code or other external element so strongly associated with the concerned individual's identity as to equate individuation. Objects and codes strongly associated to an individual will not generally be considered identifiers.	Identification does not always require individuation. Identification can be performed through identification by name, or with reference to a code or other external element strongly associated with the concerned individual's identity. Objects, codes, opinions, biological features, and other external features can be considered identifiers so long as these bear a strong, subjectively appreciated connection to the concerned natural person.
Personal character of data	Personal data must relate to the concerned identifier in content.	Personal data can relate to the concerned identifier in content, purpose, or effect.
Relevance of privacy interest	Case law is divided as to the necessity for a privacy interest to arise in data for the data to be considered personal.	No privacy interest need arise in data for it to be considered personal.
Contextual analysis of data identifiability	The identifiability analysis is contextual and is performed in the circumstances of the data's use, accounting for the serious possibility of identification inherent to the circumstances of data use. The consideration of relevant contextual elements is holistic and accounts for methods of re-identification that are illicit, inadvertent, or that have a stochastic probability of success.	The identifiability analysis is contextual and considers identifiability relative to the data controller and to a matrix of third parties sufficiently proximate to the controller in relationship or having a plausible motive to perform re-identification. Only means of performing re-identification that are reasonably likely to be used are considered. Methods of re-identification that are illicit or inadvertent appear to be excluded from the analysis. It is unclear if methods of identification that have a stochastic probability of success are accounted for.
Risk-based analysis of data identifiability	The identifiability analysis considers the "serious possibility" or "reasonable expectation" of identification.	Reasonable likelihood excludes methods of re-identification that are "practically impossible" to employ, that are prohibitively time-consuming, expensive, personnel-draining, or resource-intensive.

Figure 3: Comparative Table of Canadian and European Data Privacy Concepts