AUTHOR:
Lee Swales[1]

AFFILIATION:
[1]School of Law, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Lee Swales

EMAIL:
swalesl@ukzn.ac.za

HOW TO CITE:
Swales L. The *Protection of Personal Information Act* and data de-identification. S Afr J Sci. 2021;117(7/8), Art. #10808. https://doi.org/10.17159/sajs.2021/10808

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

KEYWORDS:
anonymisation, pseudonymisation, GDPR, personal information, POPIA

PUBLISHED:
04 June 2021

The *Protection of Personal Information Act* and data de-identification

Data have become an exceptionally valuable resource. In light of the COVID-19 public health emergency, data sharing and the concept of open science has gathered momentum.[1] The advantages and disadvantages of open science notwithstanding, a pressing issue for the scientific community to consider – particularly in relation to health research – relates to the de-identification of data, and the impact of the *Protection of Personal Information Act 4 of 2013* (POPIA) on research activities in this context. For the purposes of this Commentary, 'health research' refers to scientific research designed to learn more about human health with a view to preventing, curing and treating diseases. This type of research invariably requires the use of personal information as defined in POPIA.

On 23 September 2020, the Academy of Science of South Africa (ASSAf) announced that it would be embarking on a process to facilitate the development of a Code of Conduct for all scientific research activity with a view to submitting this Code to the Information Regulator for approval in July 2021.[2] Accordingly, the purpose of this Commentary is to: (1) discuss data de-identification and related concepts; (2) consider how data de-identification applies in the context of scientific practice in South Africa; and (3) consider relevant data de-identification principles in selected relevant foreign jurisdictions.

Background to POPIA

POPIA was the result of a painfully slow law reform process that was initiated in 2000 by the South African Law Reform Commission. The process operated under the name 'Project 124: Privacy and data protection', and, following an *issue paper* in 2003 (which announced an investigation into data protection, articulated the aim of the investigation, and pointed out solutions while also requesting comment), the project delivered a *discussion paper* in 2005 (which set out the South African Law Reform Commission's preliminary findings and recommendations and invited further comment). Thereafter, a *final report* was published in August 2009 – this report summarised the investigation, gave a detailed exposition of the applicable law, and set out draft law (known as a Bill) on protection of personal information.

POPIA was finally promulgated on 19 November 2013. Certain parts of the Act were made effective from 11 April 2014; however, the majority of the Act was effective from 1 July 2020. Critically, in terms of section 114 of POPIA, all parties have 12 months from the effective date to be fully compliant; 1 July 2021 is therefore the date by which all parties must be ready to comply with the Act.

POPIA will create a new data protection regime in South Africa, and, for the first time, the country will have a comprehensive data protection statute for all sectors – this will bring South Africa in line with many other developed nations where data protection laws are now the norm rather than the exception. The Act animates and gives effect to the right to privacy which is specifically protected by section 14 of South Africa's Constitution. Although the right is not absolute, it is now generally accepted that all persons in South Africa have a right to protection from unwanted collection and use of personal information.

However, this new regime should *not* represent a sea change for health research; treating data privately, securely and ethically should be something with which health researchers and scientists are familiar. For almost 20 years, the *National Health Act 61 of 2003* has regulated health records (see, in particular, Chapter 2 and sections 14–17 thereof which deal with confidentiality, access to records, and the protection of records). In addition, the *Health Professions Act 56 of 1974* establishes a Health Professions Council which has set out detailed ethical guidelines for good practice (see especially booklet 5 dealing with confidentiality). A thorough examination of these related provisions is beyond the scope of this Commentary, suffice to say: POPIA will not stand alone, and although it is now the point of departure when considering data protection in South Africa, depending on the context, it must be read together with other relevant legislation.

Data de-identification and POPIA

Personal information is widely defined in POPIA, and includes names, identity numbers, address information, online identifiers such as IP addresses, and, in the health research context, medical records of a patient, biometric data, and genomic data. Importantly, in terms of section 6 of POPIA, the Act will not apply to data 'de-identified to the extent that it cannot be re-identified again'. This principle, although expressed differently, is consistent with data protection legislation around the world – see, for example, Recital 26 of the European Union's Directive 95/46/EC (which is the General Data Protection Regulation also known as the GDPR)[3], and the American 'Privacy Rule' in relation to health information, articulated in section 164.514 of the *Health Insurance Portability and Accountability Act of 1996* (HIPAA).

Although the terms de-identification, anonymisation, and pseudonymisation are sometimes used interchangeably, there are subtle distinctions[4] in the meanings of these terms – and it should be noted that POPIA uses the term 'de-identification'. It is defined as:

> '***de-identify***', *in relation to personal information of a data subject, means to delete any information that—*
>
> *(a) identifies the data subject;*
>
> *(b) can be used or manipulated by a reasonably foreseeable method to identify the data subject; or*

(c) can be linked by a reasonably foreseeable method to other information that identifies the data subject,

*and '**de-identified**' has a corresponding meaning*

As a result, de-identification in terms of POPIA is a process whereby a person takes steps to delete all personal information that can identify a data subject in the data set. In the context of health research, for data to be classified as de-identified, no person within the relevant research organisation must be able to identify the data subject by considering the data set itself, and by considering other information in conjunction therewith. Therefore, using a *reasonably foreseeable method*, a person should not be able to manipulate the data to identify a data subject – for example by changing or sorting columns and/or data, or by editing the characteristics or permissions of a file to reveal information that could identify a data subject. Further, using a *reasonably foreseeable method*, a person should not be able to use other data to link to the data set to identify a data subject; for example, by using other related or unrelated data that are available either publicly or to that person specifically. In addition, if the objective behind de-identification is to ensure that POPIA does not apply to the processing of that data, section 6 of the Act places a further condition on parties – namely, that the de-identified data *cannot be re-identified again*.

This raises two questions: What is a reasonably foreseeable method? And, what does this definition – read together with section 6 – mean practically? Generally speaking, it means that the typical researcher with the usual skills, expertise and knowledge of someone working in that field, should not be able to identify a data subject in the data set. (Note that in legal terms, when the term reasonable is used, the determination is achieved objectively.) Practically, when making this determination, one must consider the data being used, the characteristics of that data, as well as other data that are available to the researcher. One must also consider section 6 which stipulates that the data should not be able to be re-identified. With large swathes of data now available publicly via the Internet, and with increasing amounts of data being shared and available electronically in many different databases, this determination can be problematic. Where a principal investigator is in doubt, it is suggested that a final determination is made by an external expert with no links to the project (a person with no potential conflict of interest, and with the necessary skills to make the determination).

In a similar vein to de-identification, anonymisation is typically defined as a process in which personal information is removed from data so that a data subject cannot be identified. The GDPR defines anonymous information in Recital 26 as 'information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable'. This definition is exceedingly similar to de-identification in POPIA, and although it may appear redundant, there are some international academics and medical professionals[5] who are of the view that these terms have distinct meanings, and that in order to foster conceptual clarity, a clear distinction should be drawn between the two terms. Briefly put, the opinion is that although de-identification removes personal information, it is still possible to re-identify the data (although it should be difficult, time-consuming and improbable that the typical researcher would be able to re-establish the link between the data and the person). However, in contrast to de-identification, anonymisation is a process whereby a researcher can practically never identify a data subject. The data are stripped so that it is virtually impossible to identify a data subject – the data are anonymised to an irreversible extent. The key difference, according to this view, is that with de-identification the process may be reversed, whereas with anonymisation it is irreversible and virtually impossible to re-identify the data. If one accepts this distinction, which is admittedly subtle, in light of section 6 and the exclusions to the Act, arguably POPIA should have rather used the term 'anonymisation' instead of 'de-identification' (given that the Act requires that data cannot be re-identified again, this appears more consistent with anonymisation than with de-identification). Alternatively, section 6 of POPIA should have been crafted on a similar basis to section 164.514 (b) of HIPAA (where data is considered de-identified if certain information is removed, or if the chance of re-identification is very low and statistically improbable). That debate notwithstanding, the correct term in South Africa is currently 'de-identification', although some authors do refer to the terms 'de-identification' and 'anonymisation' interchangeably.[6]

Another term that often features in the context of data protection is 'pseudonymisation'. Although this term is not used in POPIA, Article 4(5) of the GDPR defines pseudonymisation as a method by which personal data are processed such that the personal information can no longer be attributed to a data subject without the use of additional information, provided that the additional information is kept separately and subject to technical and organisational measures to ensure the data are not attributed to a data subject. Usually, this measure is taken as a step to ensure security of the data, to avoid bias, and to provide a level of integrity to the study. In these circumstances, someone in the organisation will have access to a master file or some other data that will facilitate the identification of the data subject if necessary (for example, the data subject may need to be identified quickly if an incidental finding is made, for audit purposes, or in the event of some medical emergency).

Scientific practice: POPIA will apply in most circumstances

In a South African context, other than the definition, POPIA does not contain any specific provision that deals with data de-identification directly. The term is mentioned in three sections of the Act (section 1, section 6, and section 14), but there is no specific guidance on how to achieve data de-identification, or any other detail in relation thereto. It is likely that after the Act has come into full effect in July 2021, the Information Regulator (the body responsible for enforcement, monitoring and education) will produce a guidance note on these issues, or that an industry Code of Conduct – such as the one being prepared by ASSAf – will articulate best practice and tips in relation thereto. For the time being, for analogous advice on techniques in relation to anonymisation, as well as useful case studies and practical examples, see the guidance set out by the United Kingdom's Information Commissioner's Office[7]. For further practical insight, see further the Singaporean Personal Data Protection Commission[8].

If the goal is to ensure that POPIA need not apply to the data in question, as noted above, researchers must ensure that all personal information that can identify a data subject is removed, and that it cannot be re-identified by anyone in the organisation. By way of example, if a data set contains no actual names or identity numbers or other personal information of a group of persons with a rare disease, but does contain birth dates or physical addresses, it is probable that another researcher could identify an individual in the study by using other data sets, and, in this instance, one would need to consider further steps in order to classify the data as de-identified (such as removing exact birth dates by giving a range, or by removing physical addresses and providing province or post code information). As a result, before data can be re-purposed or published (assuming one does not wish to need to comply with POPIA), the data must be sufficiently bereft of personal information to be considered de-identified; further, it must not be reasonably possible to reverse the process and re-identify the data.

It appears that, given the ethical imperatives and objectives of medical studies (as well as audit requirements), the data will often not be de-identified because someone in the organisation will have the ability to identify a data subject. Consequently, researchers should be cautioned against operating under the belief that POPIA does not apply to them because an individual (or large parts of the team) cannot identify a data subject – if someone (even if only one person) in the organisation has the ability to identify a data subject (via access to a master file, or using some other technique to link the data) the data will not be regarded as de-identified, and POPIA will apply. In this instance, these techniques should rather be referred to as pseudonymisation, and viewed as one of the measures taken to ensure compliance with the eight conditions of POPIA.

A foreign perspective on data de-identification

The comparable US legislation (HIPAA's Privacy Rule) seeks to protect identifiable health information; so, although it is similar in many respects

to POPIA, it applies to a certain field only (see Section 164.514 (a)–(c)). In terms hereof, information will be de-identified if it is stripped of 18 specific identifiers, *or* if it is determined by a professional statistical analyst with appropriate knowledge and experience that the risk is very small that the information, on its own, or with other information, could be used to identify a data subject. I suggest that a good rule of thumb to achieve de-identification in South Africa would be to ensure that the 18 identifiers of an individual set out in HIPAA are removed from the data set; the elimination of the 18 identifiers is known as the safe harbour method, whereas the second avenue of achieving compliance is known as the expert determination method and relies on a statistician verifying that the risk of identification of a data subject is very low. The identifiers to be removed (adapted for South Africa) are: names, addresses (except city, province and post code), all elements of dates (except a year or dates in ranges), telephone numbers, fax numbers, email addresses, identity numbers, medical record numbers, medical aid details, account numbers, certificate/licence numbers, vehicle identifiers and serial numbers, device identifiers and serial numbers, URLs, IP addresses, biometric identifiers[9,10], photographs, any other unique identifying number, characteristic or code. Once removed, and assuming the data cannot be re-identified again, one can assume that the data are de identified.

In the United Kingdom, data protection is regulated by the *Data Protection Act of 2018*, read together with the United Kingdom GDPR. The legislation is very similar to that in South Africa, and as a general proposition, the system of law operates on a similar basis. The Information Commissioner's Office (akin to South Africa's Information Regulator) has published a code of practice on anonymisation, and of relevance for present purposes is the 'motivated intruder test' it sets out therein. The test involves assessing whether a 'motivated intruder' can identify the individual in the de-identified data – it is assumed that this person is reasonably competent, has access to all publicly available information, and would employ investigative techniques; however, the 'motivated intruder' is assumed to not have any specialist skills such as hacking, and to not resort to criminality such as burglary or unauthorised use of secured data. Therefore, in borderline cases, or where one is unsure of whether data are de-identified, applying this fictitious test can assist an information officer or research team to determine whether data are de-identified. The guide also provides some useful anonymisation techniques, case studies, and practical examples – although the terminology in this context is different (anonymisation instead of de-identification), I suggest that until the Information Regulator produces something similar, and in the absence of a Code of Conduct to provide further guidance, the guidance in this UK code of practice will assist local researchers by providing valuable insight and practical examples.

Conclusion

POPIA should not be feared. The legislation marks a watershed moment in South African law and will ensure that the country keeps abreast of foreign developments. Although there are no definitive interpretations or guidance notes on data de-identification, a similar approach is used around the world in a variety of jurisdictions. It is hoped that the Information Regulator produces a code of practice or guidance note, or alternatively, that a Code of Conduct for researchers explains these issues in more detail and clarifies some of the terminology and processes.

For now, as a short-term measure, I suggest that where doubt exists, South African researchers should: (1) look to the United Kingdom's Information Commissioner's code of practice on anonymisation; (2) alternatively to point 1 (or in addition thereto), review section 164.514 (b) of HIPAA for insight on how to interpret whether data have been de-identified; and (3) follow the GDPR definition of pseudonymisation.

References

1. Organisation for Economic Cooperation and Development (OECD). Making open science a reality. OECD Science, Technology and Industry Policy Papers No. 25. Paris: OECD Publishing; 2015. http://dx.doi.org/10.1787/5jrs2f963zs1-en
2. Thaldar D, Townsend B. Protecting personal information in research: Is a code of conduct the solution? S Afr J Sci. 2021;117(3/4), Art. #9490. https://doi.org/10.17159/sajs.2021/9490
3. Roos A. The European Union's General Data Protection Regulation (GDPR) and its implications for South African data privacy law: An evaluation of selected 'Content Principles'. Comp Int Law J South Afr. 2021;53(3), Art. #7985. https://doi.org/10.25159/2522-3062/7985
4. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. J Med Internet Res. 2019;21(5), e13484. https://doi.org/10.2196/13484
5. Kushida C, Nichols D, Jadrnicek R, Miller R, Walsh J, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care. 2012;50(Suppl):S82–S101. https://doi.org/10.1097/MLR.0b013e3182585355
6. Burns Y, Burger-Smidt A. A commentary on the Protection of Personal Information Act. Durban: LexisNexis South Africa; 2018. p. 102–103.
7. Information Commissioner's Office. Anonymisation: Managing data protection risk code of practice [document on the Internet]. c2012 [cited 2021 May 25]. Available from: https://ico.org.uk/media/1061/anonymisation-code.pdf
8. Singaporean Personal Data Protection Commission. Guide to basic data anonymisation techniques [webpage on the Internet]. c2018 [cited 2021 May 25]. Available from: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf?la=en
9. Carter A. Considerations for genomic data privacy and security when working in the cloud. J Mol Diagn. 2019;21(4):542–552. https://doi.org/10.1016/j.jmoldx.2018.07.009
10. Dankar F, Ptitsyn A, Dankar S. The development of large-scale de-identified biomedical databases in the age of genomics-principles and challenges. Hum Genomics. 2018;12(1):19. https://doi.org/10.1186/s40246-018-0147-5