Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review

Chevrier, Raphaël; Foufi, Vasiliki; Gaudet-Blavignac, Christophe; Robert, Arnaud; Lovis, Christian

How to cite

CHEVRIER, Raphaël et al. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. In: Journal of Medical Internet Research, 2019, vol. 21, n° 5, p. e13484. doi: 10.2196/13484

This publication URL: https://archive-ouverte.unige.ch/unige:136227
Publication DOI: 10.2196/13484

Review

Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review

Raphaël Chevrier[1,2], MD; Vasiliki Foufi[1,2], PhD; Christophe Gaudet-Blavignac[1,2], BSc, MD; Arnaud Robert[1,2], MSc; Christian Lovis[1,2], MD, MPH, FACMI

[1]Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland

[2]Faculty of Medicine, University of Geneva, Geneva, Switzerland

Corresponding Author:
Raphaël Chevrier, MD
Division of Medical Information Sciences
University Hospitals of Geneva
Rue Gabrielle-Perret-Gentil, 4
Geneva, 1205
Switzerland
Phone: 41 223790847
Email: raphael.chevrier@hcuge.ch

Abstract

Background: The secondary use of health data is central to biomedical research in the era of data science and precision medicine. National and international initiatives, such as the Global Open Findable, Accessible, Interoperable, and Reusable (GO FAIR) initiative, are supporting this approach in different ways (eg, making the sharing of research data mandatory or improving the legal and ethical frameworks). Preserving patients' privacy is crucial in this context. De-identification and anonymization are the two most common terms used to refer to the technical approaches that protect privacy and facilitate the secondary use of health data. However, it is difficult to find a consensus on the definitions of the concepts or on the reliability of the techniques used to apply them. A comprehensive review is needed to better understand the domain, its capabilities, its challenges, and the ratio of risk between the data subjects' privacy on one side, and the benefit of scientific advances on the other.

Objective: This work aims at better understanding how the research community comprehends and defines the concepts of de-identification and anonymization. A rich overview should also provide insights into the use and reliability of the methods. Six aspects will be studied: (1) terminology and definitions, (2) backgrounds and places of work of the researchers, (3) reasons for anonymizing or de-identifying health data, (4) limitations of the techniques, (5) legal and ethical aspects, and (6) recommendations of the researchers.

Methods: Based on a scoping review protocol designed a priori, MEDLINE was searched for publications discussing de-identification or anonymization and published between 2007 and 2017. The search was restricted to MEDLINE to focus on the life sciences community. The screening process was performed by two reviewers independently.

Results: After searching 7972 records that matched at least one search term, 135 publications were screened and 60 full-text articles were included. (1) Terminology: Definitions of the terms de-identification and anonymization were provided in less than half of the articles (29/60, 48%). When both terms were used (41/60, 68%), their meanings divided the authors into two equal groups (19/60, 32%, each) with opposed views. The remaining articles (3/60, 5%) were equivocal. (2) Backgrounds and locations: Research groups were based predominantly in North America (31/60, 52%) and in the European Union (22/60, 37%). The authors came from 19 different domains; computer science (91/248, 36.7%), biomedical informatics (47/248, 19.0%), and medicine (38/248, 15.3%) were the most prevalent ones. (3) Purpose: The main reason declared for applying these techniques is to facilitate biomedical research. (4) Limitations: Progress is made on specific techniques but, overall, limitations remain numerous. (5) Legal and ethical aspects: Differences exist between nations in the definitions, approaches, and legal practices. (6) Recommendations: The combination of organizational, legal, ethical, and technical approaches is necessary to protect health data.

Conclusions: Interest is growing for privacy-enhancing techniques in the life sciences community. This interest crosses scientific boundaries, involving primarily computer science, biomedical informatics, and medicine. The variability observed in the use of the terms de-identification and anonymization emphasizes the need for clearer definitions as well as for better education and dissemination of information on the subject. The same observation applies to the methods. Several legislations, such as the

American Health Insurance Portability and Accountability Act (HIPAA) and the European General Data Protection Regulation (GDPR), regulate the domain. Using the definitions they provide could help address the variable use of these two concepts in the research community.

(J Med Internet Res 2019;21(5):e13484) doi: 10.2196/13484

KEYWORDS

anonymization; anonymisation; de-identification; deidentification; pseudonymization; privacy; confidentiality; secondary use; data protection; scoping review

Introduction

Background

In 2003, the National Institutes of Health (NIH) released its final statement on sharing research data. The NIH made the provision of a data-sharing plan mandatory for any funding starting at US $500,000 per year [1]. This statement, among other published work [2-5], accelerated the sharing of research data worldwide in parallel to the growing availability of data and information technologies. In this context, the research community gained an unprecedented capacity to access and analyze large amounts of health data, originating partly from nonresearch sources. The use of medical data for a different purpose than the one it was initially collected for is commonly called "secondary use of medical data" [3]. This particular use of health data is subject to technical and semantic problems as well as legal, ethical, and societal concerns. To comply with the legal and ethical principles, researchers have two main options to access and use medical data for a secondary purpose [6]. One option is to gain patients' consent specifically for the new purpose of their research. This is generally complicated and costly [7]. Alternatively, they can de-identify the data, since the law permits the disclosure of clinical information if it has been correctly de-identified [8]. Institutional review boards (IRBs) generally waive the need for consent in this situation [9]. The existence of the second option gives de-identification and anonymization a pivotal role in biomedical research. Consequently, the availability of reliable techniques to protect privacy becomes essential for the research community to leverage the secondary use of medical data [10].

Despite all efforts, an important gap still exists between the needs and the access to massive data in science. Large collaborative data-sharing projects are somehow below expectations and the research community is calling for improved open data and open science [11]. Some authors have proposed explanations as to why data sharing is more complicated in practice than in theory [3]. An article has considered the influence of policies and of our capacity to protect the data on our ability to share it [12]. Reviews have been published on the techniques and systems aiming at protecting health data privacy [13,14]; one has collected and studied the known re-identification attacks on health data [15], and another has looked specifically into the security and privacy issues related to electronic health records [16]. Various techniques aim at protecting the medical data subjects' privacy. Those that do not strictly represent an anonymization or de-identification process are not part of the scope of this review. Cryptography, privacy-preserving record linkage [17], and differential privacy [18] are among these techniques.

Although advanced probabilistic privacy-enhancing methods have been studied and applied for over three decades in other areas [19], their application to medical data is a fairly recent interest for the biomedical research community. A striking example is the late introduction of *data anonymization* (2016) and other central concepts of health data privacy (eg, *personally identifiable information*) in the Medical Subject Headings (MeSH) thesaurus of the US National Library of Medicine. Over the last few years, a great amount of expert literature was produced on anonymization and de-identification techniques for medical data. However, publications providing the readers with a broad understanding of these techniques, and addressing their application in life sciences and clinical research in a comprehensive way, are lacking. As a consequence, the fundamental concepts remain either unknown to the research community or difficult to comprehend. Adding to the confusion, a well-documented and long-standing ambiguity exists in the vocabulary used by those who contribute to the practice [20,21]. In particular, the terms *de-identification* and *anonymization* have been used with different meanings by researchers. De-identification is frequently, but not exclusively, used in the biomedical literature to refer to rule-based techniques. These techniques often apply the rules provided in the *Safe Harbor* method of the American legislation (ie, the Health Insurance Portability and Accountability Act [HIPAA]). On the other hand, anonymization is commonly, but not exclusively, used in the biomedical literature to refer to statistical or probabilistic techniques. Turning to the legislations to clarify the meaning of these terms can bring further confusion. Although researchers tend to use two different terms—de-identification and anonymization—to refer to one approach or the other, the American law itself regards both approaches (ie, rule-based or probabilistic) as ways to achieve de-identification. The first follows the *Safe Harbor* method—§164.514(b)(2)—and the second follows the *Expert Determination* method—§164.514(b)(1). The European legislation (ie, the General Data Protection Regulation [GDPR]), on the other hand, does not use either of the terms.

A growing number of health care data breaches are being reported [22], some resulting directly from a failure to anonymize or de-identify the data properly [23]. In this context, it seems essential to review the literature published on this rapidly evolving domain to inform researchers, doctors, lawmakers, and the public about instruments that are becoming indispensable to researchers. This is true, especially since these instruments have bearing on subjects of paramount importance

for the future of medical research, namely, data sharing, data privacy, and public trust in health care research and institutions.

Objectives

The aim of this work is to better understand how the life sciences research community defines, comprehends, and uses the concepts of de-identification and anonymization. Providing a broad perspective on the field, this review should also contextualize these concepts and their application in today's biomedical research domain. To attain these goals, the reviewers identified six key aspects to study, which are presented in Textbox 1. These aspects are central to this work; they guided the data collection and they structured the Results section.

Textbox 1. Objectives: subjects of focus for this scoping review.

1. Vocabulary, definitions, and understandings of the terms *anonymization* and *de-identification*.
2. Authors' backgrounds and places of work.
3. Reasons for anonymizing or de-identifying health data.
4. Limitations of anonymization and de-identification techniques.
5. Legal and ethical implications of the practice.
6. Experts' recommendations.

Methods

Overview

Scoping reviews represent an increasingly popular type of review [24], which allows for the mapping of concepts in a field of interest. They are intended to study complex and overlapping domains, particularly when they have not been reviewed comprehensively before [25]. To conduct this work, the authors used the guidance proposed by the Joanna Briggs Institute for the conduct of scoping reviews [26].

The first step of the scoping review process was to perform a preliminary, nonsystematic survey of the literature regarding de-identification and anonymization. This survey identified the key concepts, the concerns, the challenges, and the gaps in the domain. This information was used to define the study's objectives and to design the study protocol.

Article Identification and Selection

Search Strategy

Aiming to focus this work on the life sciences researchers' community, the articles were sourced selectively from one database: MEDLINE [27]. To maximize the sensitivity and specificity of the search query, several strategies were tested and implemented. The terms "de-identification" and "anonymization"; their lexical variants (eg, "de-identif*," and "anonymi*"); and their spelling variants (eg, "deidentified" without hyphen and "anonymisation" in British English) were used. Alternative spellings were proven effective in a previous literature review on the same topic [28]. Numerous candidate terms were tested, here are some examples: "privacy protection," "data protection," "confidentiality," "personal data," "medical data," "re-identification," and "breaches". None of these terms increased the sensitivity of the search compared to the terms "de-identification," "anonymization," and their variants. The same conclusion was reached regarding the use of the MeSH-controlled vocabulary. Finally, search-field descriptors—[ti] (Title) and [tiab] (Title/Abstract)—were used, and the terms were combined between themselves using Boolean operators. A full description of the search query is provided in the Results section.

Inclusion Criteria

This work analyses the literature published between November 1, 2007, and November 1, 2017. Only original research articles and review articles available in full text through the University of Geneva's library network were considered. Additionally, publications had to meet at least one of the following three criteria to be included:

1. The subject of the article is the process of rendering medical data as less identifiable using computer techniques (ie, de-identification or anonymization).
2. The article focuses on sharing medical data; however, protecting the patients' privacy using computer techniques is also discussed.
3. The article presents legal and ethical aspects of sharing medical data, and the concept of de-identification or anonymization is discussed.

Exclusion Criteria

The literature addressing certain data types was excluded: video recordings, photographic images, radiological images, and geolocation data. This decision was made on the basis of the information found during the preliminary literature survey [28,29] and was confirmed after discussions with experts. Short reports, posters, and editorials were also excluded.

Data Collection

Based on the list of six objectives (see Textbox 1), information categories were defined (see Table 1). Quantitative and qualitative data were collected from the articles. Quantitative information was extracted for certain categories and statistical analysis was performed on this data. Qualitative information was collected for the categories not suited to quantitative analysis. This second approach was nonetheless important, as it allowed us to bring together the views of some experts and to identify consensus or disagreements.

Table 1. Categories of information used to collect quantitative and qualitative data from the reviewed articles.

Type of data	Categories of information
Quantitative	• Journal • Year of publication • Author(s) • Authors' backgrounds • Authors' places of work • Presence of the terms "de-identification" and "anonymization" • Definitions of the terms "de-identification" and "anonymization" • Meanings given to the terms "de-identification" and "anonymization"
Qualitative	• Purposes of de-identification and anonymization • Limitations of the privacy-enhancing techniques • Ethical or legal considerations • Suggestions and recommendations • Data utility and information loss • Data sharing in biomedical research • Types of data subjected to anonymization or de-identification • Public opinion on privacy-enhancing techniques and health data sharing

To determine the backgrounds of the authors, points were attributed to domains (medicine, computer science, law, etc) according to each author's professional affiliation and academic qualifications. Up to three authors were included per publication (ie, first, second, and last author), based on a previous research study, which showed that the most significant contributions were made by these authors [30]. All publications included in the review were considered. The information about the authors was collected manually from the articles, from the authors' or organizations' websites, and from other sources, such as Google Scholar, Open Researcher and Contributor ID (ORCID), ResearchGate, etc.

Results

Study Selection and Characteristics

The literature search retrieved 135 articles from the sizeable number of existing records—7972 after the removal of duplicates—containing at least one of the search terms used. The breakdown of the search query shows the number of records at each level (see Figure 1).

The search query identified 135 records in MEDLINE corresponding to the keyword search; the records were then manually screened according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology (see Figure 2). Among them, 103 records were in the considered time frame. Three records were excluded because the full text could not be retrieved. An additional 40 articles were excluded based on the focus of the paper, the data type considered, or the publication type. During this process, five records raised questions about their potential eligibility. A third reviewer was involved to reach consensus.

Figure 1. Architecture and breakdown of the search query with the number of records at each level. [ti]: Title; [tiab]: Title/Abstract.

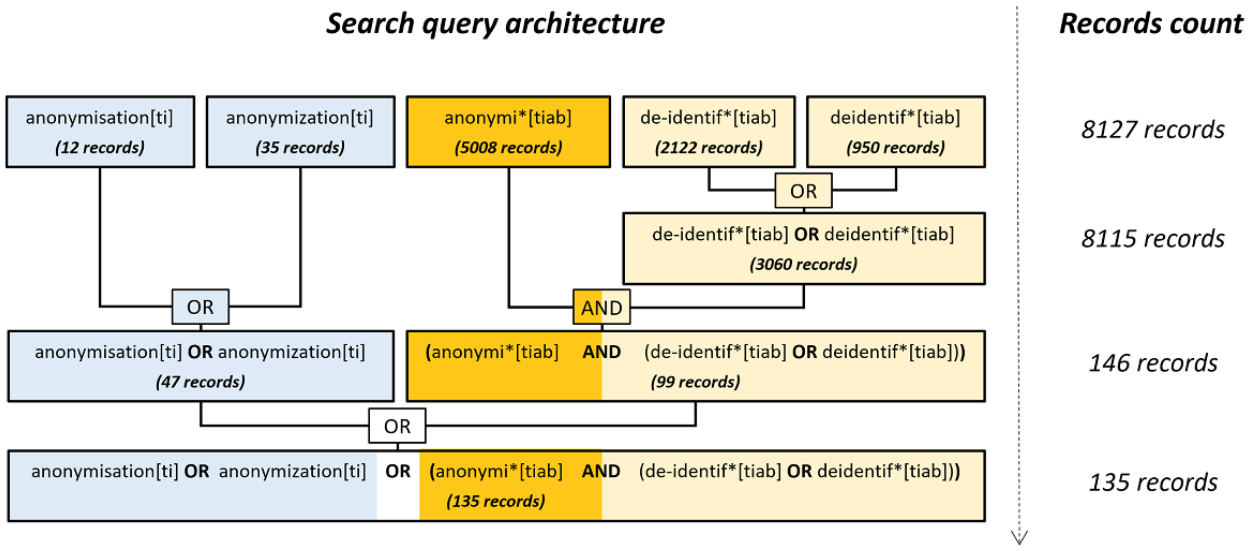
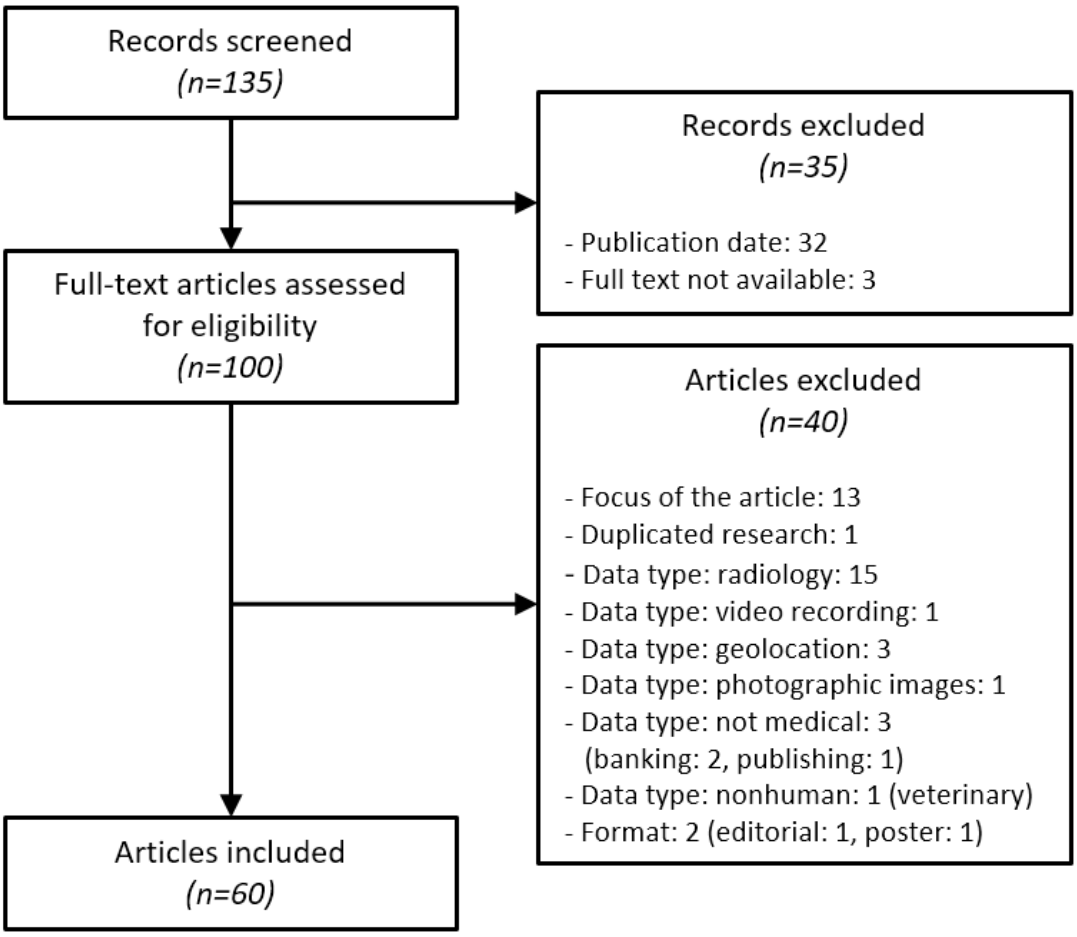


Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for the scoping review process (ie, screening, eligibility, and inclusion).



The process resulted in the inclusion of 60 articles; the list is available in Multimedia Appendix 1. The 60 articles came from 32 different scientific journals. Corrected for the five journals that did not have a registered impact factor in 2017, the average impact factor of the journals included in this review was 2.859, ranging from 9.504 for the highest to 0.304 for the lowest, with a median of 2.766. More than a third of the articles (23/60, 38%) were published after 2015 (see Table 2).

Table 2. Characteristics of the 60 articles included in the review and of the journals where they were published.

Characteristics	Count (N=60), n (%)
Year of publication	
2008-2009	5 (8)
2010-2011	11 (18)
2012-2013	10 (17)
2014-2015	11 (18)
2016-2017	23 (38)
Scientific domains of the journals	
Biomedical informatics	32 (53)
Engineering	8 (13)
Public health, methodology, and epidemiology	6 (10)
Bioethics and law & health policies	5 (8)
Medicine: biomedical sciences	5 (8)
Medicine: clinical	4 (7)

Vocabulary and Definitions

Half of the articles (29/60, 48%) provided a definition of de-identification or anonymization (see Table 3).

The attempts at defining the terms were rare, and the definitions often vague, inconsistent, or even contradictory (see Table 4). Referring to the HIPAA *Safe Harbor* method of de-identification, one article correctly recommended the removal of 18 types of protected health information (PHI) [13]. Another suggested the removal of 17 types of PHI [31]. Regarding the processing of the types of PHI, one article proposed to "hide" or "remove" them [13], while another suggested to "extract" or "replace" them with pseudonyms [32]. Concerning anonymization, the variability was similar. One article presented the process as the removal of the patients' names [33]. Another considered it a much more radical alteration of the data, which would be virtually impossible to reverse [28].

Conflicting representations of de-identification and anonymization were uncovered (see Textbox 2). In some articles, the terms are used interchangeably to refer to the same concept [34-37], while in others they outline strictly different processes [13,19,28,38].

The researchers' representations of de-identification and anonymization, as similar or different concepts, were counted from the reviewed articles to determine whether or not there was a consensus among the experts. The results are presented in Table 5. The 38 authors who used both terms were evenly split between those who considered the two notions to be identical (19/60, 32%) and those who considered them to be different (19/60, 32%).

The 19 researchers who only used or discussed one concept in the core of their articles mentioned the second one in the keywords or title. From the reviewers' perspective, this finding reinforces the idea that de-identification and anonymization are synonyms in many people's minds.

Table 3. Presence of definitions for the terms *de-identification* or *anonymization* in the reviewed articles.

Terms with definitions	Count (N=60), n (%)
De-identification	26 (43)
Anonymization	12 (20)
Both	9 (15)
None	31 (52)

Table 4. Examples of attempts to define the terms *de-identification* or *anonymization*.

Terms	Definitions
De-identification	"For clinical data to be considered de-identified, the HIPAA 'Safe Harbor' technique requires 18 data elements (called PHI: Protected Health Information) to be removed...de-identification only means that explicit identifiers are hidden or removed." [13]
	"Under Safe Harbor, data are considered de-identified if 17 listed types of identifiers are removed." [31]
	"de-identification where explicit identifiers (e.g., Protected Health Information [PHI] elements) are extracted or replaced with 'pseudonyms'" [32]
	"De-identification of medical record data refers to the removal or replacement of personal identifiers so that it would be difficult to reestablish a link between the individual and his or her data. Although a de-identified dataset may contain an encrypted patient identifier with which authorized individuals could relink a patient with his or her dataset, this dataset must not contain data that will allow an unauthorized individual to infer a patient's identity from the existing data elements." [28]
Anonymization	"The anonymization consists in removing the patients' names from the records: unfortunately, other pieces of information enable to identify the patients." [33]
	"anonymization implies that the data cannot be linked to identify the patient" [13]
	"the process of rendering data into a form which does not identify individuals and where identification is not likely to take place" [10]
	"Data anonymization is the process of conditioning a dataset such that no sensitive information can be learned about any specific individual." [19]
	"Anonymization refers to the irreversible removal of the link between the individual and his or her medical record data to the degree that it would be virtually impossible to reestablish the link." [28]

Textbox 2. Discrepancies in understanding and using de-identification and anonymization in relation to each other.

Anonymization = de-identification:

- "Access to de-identified (anonymized) health records would in many cases be sufficient." [34]
- "Anonymization: Redaction, perturbation, or generalization of those attributes that could be used, alone or in combination, to associate a given record with a specific person. Also called 'de-identification.'" [35]
- "Recent renewed interest in de-identification (also known as 'anonymisation') has led to the development of a series of systems in the United States with very good performance on challenge test sets." [36]
- "As has been seen, the European regime for privacy does not require the de-identification (anonymization) of personal data used in genomic databases or biobanks." [37]

Anonymization ≠ de-identification:

- "we note that a recent analysis of matching attacks against a large, public, de-identified (although not anonymized) dataset independently came up" [19]
- "Anonymization and de-identification are often used interchangeably, but de-identification only means that explicit identifiers are hidden or removed, while anonymization implies that the data cannot be linked to identify the patient (i.e. de-identified is often far from anonymous)." [13]
- "De-identification of medical record data refers to the removal or replacement of personal identifiers so that it would be difficult to reestablish a link between the individual and his or her data...Anonymization refers to the irreversible removal of the link between the individual and his or her medical record data to the degree that it would be virtually impossible to reestablish the link." [28]
- "The term 'anonymization' is not identical to 'de-identification.' De-identification is the removal of attributes known to increase the risk of identification, and this can be seen as a preliminary step for producing anonymous data. It requires, however, a further assessment as to whether the de-identification process achieves anonymization." [38]

Table 5. Researchers' understanding of de-identification and anonymization as similar or different concepts.

Use of the terms in the articles	Count (N=60), n (%)
Only use or discuss one concept	19 (32)
De-identification and anonymization are two different concepts	19 (32)
De-identification and anonymization are used interchangeably	19 (32)
Ambiguous with regard to the meaning of both terms	3 (5)

Authors' Backgrounds and Places of Work

Applying the scoring system presented in the Methods section, we counted 163 authors for the 60 publications. A total of 248 background points were attributed to 19 different domains (see Table 6).

The first seven fields represent 90% of the researchers' backgrounds. On average, one researcher was awarded 1.52 research field points. A total of 14 researchers published more than one article (ie, 2-8 articles). Out of 14 prolific authors, 13 (93%) had a background in the three leading domains. Removing the duplicates revealed 121 unique authors. The number of domains and their ranking remained unchanged with and without duplicates, with a slightly smaller gap between the first three domains and the others when duplicates were removed. The background of 7 authors could not be found; this represents a margin of error of 4.3%.

Regarding the place of work, the United States was the largest contributor with 25 articles (25/60, 42%), followed by Germany, the United Kingdom, and Canada combined (23/60, 38%). The predominance of publications from the US-based research groups is noticeable particularly between 2010 and 2012. After this period, their contribution decreases in absolute number and, more importantly, in relation to other groups, due to the arrival of new groups from 2014 and the rapid growth of publications on the topic of anonymization and de-identification. As a result, the leading position that American researchers (ie, Canada and the United States) held until 2013 was caught up to by researchers from other countries in 2014. Since 2015, European groups have been publishing an equal or greater number of articles than the Americans on this topic (see Figure 3).

Purpose of Anonymization and De-Identification

Most often, the authors mentioned the secondary use of medical data without specification as to the purpose of their research. When specified, their objective was to enable and support biomedical research [7,32,39-41]. Regarding the research domains, *genetics and genomics* [42-49] were the most frequently cited, followed by *personalized health and precision medicine* [48,50-52]. Improvement in the domains of *epidemiology* and *public health surveillance and reporting* were among other anticipated benefits of developing privacy protection techniques [8,53]. The protection of privacy was implicit in most projects but was also explicitly cited as a standalone objective in some publications [50,54]. Complying with regulations and policies was a motivation expressed by certain authors [46,55]. Several other reasons were found, as shown in Textbox 3.

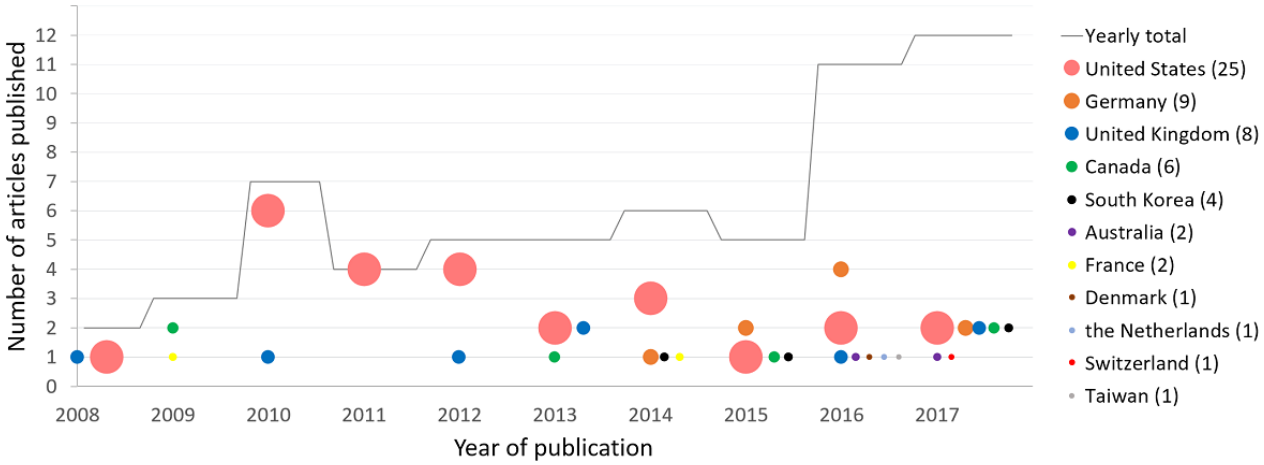
Table 6. Background points awarded to the authors of the reviewed articles. The authors are separated by authorship position: first, second, and last.

Research field	First author (N=92), n (%)	Second author (N=72), n (%)	Last author (N=84), n (%)	Total count (N=248), n (%)
Computer science	36 (14)	26 (10)	29 (12)	91 (36.7)
Biomedical informatics	16 (6)	15 (6)	16 (6)	47 (19.0)
Medicine (MD[a])	13 (5)	9 (4)	16 (6)	38 (15.3)
Epidemiology and statistics	6 (2)	3 (1)	7 (3)	16 (6.5)
Mathematics and biomathematics	6 (2)	5 (2)	5 (2)	16 (6.5)
Law	3 (1)	3 (1)	2 (1)	8 (3.2)
Psychology	2 (1)	3 (1)	2 (1)	7 (2.8)
Linguistics	2 (1)	0 (0)	2 (1)	4 (1.6)
Project management	1 (0)	1 (0)	1 (0)	3 (1.2)
Bioethics and humanities	1 (0)	2 (1)	0 (0)	3 (1.2)
Public health	1 (0)	0 (0)	1 (0)	2 (0.8)
Neuroscience	2 (1)	0 (0)	0 (0)	2 (0.8)
Behavioral economy	0 (0)	2 (1)	0 (0)	2 (0.8)
Journalism	1 (0)	1 (0)	0 (0)	2 (0.8)
Biology and microbiology	1 (0)	0 (0)	1 (0)	2 (0.8)
Physics	1 (0)	1 (0)	0 (0)	2 (0.8)
Health care administration	0 (0)	0 (0)	1 (0)	1 (0.4)
Ecology and evolution	0 (0)	1 (0)	0 (0)	1 (0.4)
Business (MBA[b])	0 (0)	0 (0)	1 (0)	1 (0.4)

[a]MD: Doctor of Medicine.

[b]MBA: Master of Business Administration.

Figure 3. Representation of the 60 publications according to the date of publication, the number of articles per year, and the authors' locations. The size of the discs used on the graph represents each country's contribution in number of articles over the studied period (10 years). The exact count is shown between brackets next to each country's name.



Textbox 3. Additional reasons expressed by experts for de-identifying or anonymizing health data.

1. Publication in biomedical journals [56].
2. Teaching [34].
3. Spontaneous reporting systems to collect adverse drug events [57].
4. Limiting the administrative burden of consent in research [7,38].
5. Facilitating clinical trial data publication [35].
6. Facilitating population screening programs [58].
7. Enabling the creation of medical text corpora for natural language processing (NLP) research and development [37].
8. Protecting particularly sensitive information (eg, mental health data) [59].
9. Producing reports on prescribing patterns and drug utilization and to perform economic studies [60].
10. Performing comparative effectiveness studies [45].

Limitations of Anonymization and De-Identification Techniques

Technical and Operational Limitations

Anonymization and de-identification are time-consuming tasks, particularly when textual data is concerned [61]. The necessity for manual intervention is seen as a weakness that leaves room for human error and contributes to lengthening the procedure [62]. The difficulty in generalizing and scaling the de-identification and anonymization procedures, as well as the absence of broadly accepted metrics to judge their results, are recurrent concerns raised in publications [8,13,38,39,60]. The complexity of these procedures depends on the type of information involved. Structured information (eg, tabular data) is generally easier to process than unstructured information (eg, textual data) [29]. Specific types of information (eg, diagnoses of rare diseases [60]) are more identifying than others. Some types are even considered identifying by nature (eg, large genome sequencing) and presumably impossible to render anonymous [38,63]. More generally, balancing the probability of re-identification with the amount of distortion applied to the data is seen as a challenge [7,59]. Unable to overcome the interdependence between data quality and data identifiability, one has to be compromised for the other: "no existing anonymization algorithm provides both perfect privacy protection and perfect analytic utility" [19]. The re-identification risk depends on the availability of additional information. Using data linkage techniques, the presence of individuals in the protected dataset can be revealed and their personal information re-identified [51,64,65]. Because the amount of information available for comparison can only be estimated, the re-identification risk will always remain an estimate [66]. Additionally, this risk will increase over time [13,48]. These inherent weaknesses have led some researchers to express doubts about the reliability of anonymization or de-identification techniques [35,66,67].

Limitations in Accessibility and Governance

The substantial cost and the limited access to trained professionals are seen as hindrances for institutions wanting to share their data [29,33,66]. Disparities in the availability of anonymization and de-identification systems between English-speaking countries and the rest of the world is expressed by certain authors [33,40]. Textual data is primarily concerned by this problem with a critical need for natural language processing (NLP) systems in varied languages [36,68]. Authors report the lack of practical guidelines and training to assist the researchers [31,69]. They also report an absence of a consensus "regarding the effective governance of secondary research uses, beyond adherence to the terms of informed consent" [70]. Finally, several researchers point out the confusion affecting the terminology as a flaw in itself, increasing the risk of re-identification through misconceptions and misunderstandings [38,71].

Legal and Ethical Implications

General

Privacy laws and regulations provide the legal framework for the collection, processing, and sharing of personal data [71]. Differences exist between nations in the definitions, approaches, and legal practices [66]. Commonly, legal experts agree that relying on legislation alone to protect privacy would be an error [71]. Legislations are effective when used in conjunction with ethical principles, commitments in data use agreements (DUAs), and technical safeguards provided by the de-identification and anonymization process [52,72]. Stricter DUAs can be used to mitigate the loss of data quality that would otherwise be required if the technical process alone had to guarantee the privacy [48,73]. Current rules and regulations are seen by some authors as too soft to discourage attempts at re-identifying data, however, the same authors recommend consistency over severity in prosecuting the misuse of health data [71].

Accountability

The legal responsibilities and the ethical obligations are shared by all those involved in the collection or in the use of the data (ie, institutions or individuals) [10]. Research participants generally believe that anyone who uses their information, regardless of when and under which circumstances, share these responsibilities [50].

Institutional Review Boards

Review boards play an important role in the secondary use of health data. Although de-identified or anonymized data, in some cases, are not considered individually identifiable health

information, research projects involving such data generally require IRB submission and approval. In these situations, the IRB assesses the information the subjects received, what they consented to, and whether the proposed research could be conflicting with their interests [31,72]. Eventually, if the IRB approves the project in question, it waives the obligation for informed consent [7,13,31].

Experts' Recommendations

The highest level of protection can only be provided by multidisciplinary approaches combining organizational, legal, ethical, and technical safeguards [10,59,72-74]. Relying exclusively on one of these aspects would be a mistake [75]. More information and training should be provided to researchers about privacy protection and about the risks associated with data sharing [60,69,74]. Numerous researchers express the necessity to review and update the current legal framework [10,31,56,59,71]. Many authors consider the ambiguity of the vocabulary and the misuse of terms as a problem that urgently requires a cooperative effort from the expert community [19,38,56,74]. When applying anonymization techniques, researchers generally recommend favoring privacy over data quality in the process of de-identification and anonymization [75,76].

Discussion

Principal Findings

The development and the application of privacy-enhancing techniques to health data has come to represent a research domain in its own right. This domain is growing rapidly, as demonstrated by the increasing number of publications and the arrival and geographical spread of new research groups. Researchers come from different disciplines and often have qualifications in several fields themselves. Computer science, biomedical informatics, and medicine are the most prevalent backgrounds overall; the main purpose driving the development and application of privacy-enhancing techniques to medical data is to facilitate biomedical research.

At the beginning of the 2000s, great hopes leaned on our abilities to develop technical safeguards that would unleash the potential of the secondary use of medical data. Almost 20 years later, our knowledge and competences have significantly improved, although every advance has come with new interrogations and challenges. Methods are still difficult to generalize or scale and inevitably alter the data quality, which can notably hinder its use for research. A successful exercise lessens the risk of re-identification while maintaining a sufficient level of data quality for research to be performed. In this aim, legal and contractual safeguards are essential and their use can be tailored (ie, made stricter or more lenient) to each situation to mitigate the technical limitations. The research community emphasizes that the different approaches (ie, organizational, legal, ethical, and technical) are complementary and necessary to provide an acceptable level of protection. What is an acceptable level of protection, however, is not easily defined. It varies both in the views of different experts and in the legislations of the different countries.

This work confirms and further illustrates the existence of a disconcerting confusion in the domain's vocabulary affecting the understanding of the concepts at multiple levels. The vagueness and lack of consensus among the experts is worrying and requires actions. The life sciences research community is aware of this situation and is calling for clear and standardized definitions and for cross-border regulatory frameworks.

Propositions

Clear Definitions

Appropriate use of the terms *de-identification* and *anonymization* should be promoted and incentivized. As a first step in this direction, the authors of this work suggest that future publications on the subject include definitions or state which definitions are referred to. Although not universal, clear definitions are provided in two major legislations on personal data protection (ie, the GDPR and the HIPAA) and should be used where appropriate.

The GDPR defines anonymous information as "an 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" [77]. It is an irreversible state. Accordingly, the term *anonymous* should not be used to describe the process of rendering data less identifiable, which is the prevailing representation of de-identification and anonymization in the biomedical literature. To refer to the concept of rendering data less identifiable, or to the techniques that are used in this aim, the GDPR defines the term *pseudonymization*: "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

Finally, the term *de-identification* comes from the American legislation where definitions are provided: HIPAA §164.514(a) and (b). Authors using this term in their publications should refer to these definitions.

Development of Clear Guidelines

In a manner that already exists in most clinical disciplines, international guidelines regarding privacy protection should be developed, agreed upon, and made widely available to the stakeholders in the field of biomedical research. These guidelines should clarify the concepts, the definitions, and the techniques, as well as their results and risks.

Improved Dissemination and Education

A striking result of this work is the lack of information dissemination and education at all levels. It is critical that the research community gains access to the appropriate information, definitions, and guidelines on the subjects of data privacy and data protection. The public and the media should benefit from this improved access and understanding. Building trust is essential for life sciences research to leverage today's technological capabilities in accessing, sharing, and analyzing

data. With this aim, information dissemination is key (see Textbox 4).

Limitations of This Work

There are several limitations to this work. As for any scoping review based on free-text searches, contributions may have been missed despite having maximized the search sensitivity. Privacy protection of health data is a rapidly evolving domain. Between the end of the scoping review and January 2019, 14 additional publications would have to be assessed for eligibility (ie, 114 vs 100).

The fact that the literature search was limited to MEDLINE introduces a strong but deliberate selection bias toward the domain of life sciences. Within life sciences, *genomics*, *personalized health*, and *precision medicine* may be overrepresented due to their growing popularity in recent years and their characteristic need for large amounts of sensitive data.

During the data collection, it was not possible to find the background of 7 authors. This represents 4.3% of the total author count (N=163), which should not impact the validity of the results.

Conclusions

Health data is increasingly produced and used. This wealth of information should not be left dormant as it represents a real potential to fuel research and improve medicine. Multidisciplinary safeguards (ie, ethical, organizational, legal, and technical) are required to guarantee the privacy of health data subjected to secondary use. Creating an overall trusted environment to leverage scientific research in life sciences is essential. It requires building on safe and strong foundations, to have processes and structures in place to enforce these foundations, and to communicate widely with the public and the media.

Textbox 4. Recommendations for future work.

- Future publications should include definitions or state which definitions are referred to.
- Existing definitions proposed by major legislations (ie, the Health Insurance Portability and Accountability Act [HIPAA] and the General Data Protection Regulation [GDPR]) should be used where applicable.
- Global and specific guidelines should be developed to define the field of application, the process, the expected results, and the risk of the different technical approaches to privacy protection.
- Information dissemination and education should be improved across the research community for all stakeholders.

Acknowledgments

We would like to thank W Martin, LLM, for his precious help in better understanding some legal concepts.

Authors' Contributions

RC performed the preliminary literature survey; codesigned the study protocol; performed the screening and inclusion process, data collection, and data analysis; and wrote the manuscript. VF performed the screening and inclusion process and data collection and helped in writing the manuscript. CGB participated in the discussions concerning the eligibility of some articles and helped with the interpretation of the results. AR participated in the data analysis and created the figures. CL codesigned the study protocol and participated in the data analysis and the redaction of the manuscript. The manuscript has been reviewed and approved by all authors.

Conflicts of Interest

CL is Editor-in-Chief for JMIR Medical Informatics.

Multimedia Appendix 1

List of the 60 articles reviewed in this work.

[PDF File (Adobe PDF File), 500KB-Multimedia Appendix 1]

References

1. Final NIH Statement on Sharing Research Data. Bethesda, MD: National Institutes of Health; 2003 Feb 26. URL:https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [accessed 2019-01-24] [WebCite Cache ID 75fHDUrPp]
2. Institute of Medicine (IOM). Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: The National Academies Press; 2015.
3. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
4. Pisani E, Whitworth J, Zaba B, Abou-Zahr C. Time for fair trade in research data. Lancet 2010 Feb 27;375(9716):703-705. [doi: 10.1016/S0140-6736(09)61486-0] [Medline: 19913902]

5. Dukes P, Clement-Stoneham G. Data Sharing Policy. Version 2.2. London, UK: Medical Research Council; 2016 Sep. URL:https://www.mrc.ac.uk/documents/pdf/mrc-data-sharing-policy/ [accessed 2018-02-13] [WebCite Cache ID 6xCclmaMF]
6. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015 Mar 20;350:h1139 [FREE Full text] [doi: 10.1136/bmj.h1139] [Medline: 25794882]
7. El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, et al. A globally optimal k-anonymity method for the de-identification of health data. J Am Med Inform Assoc 2009;16(5):670-682 [FREE Full text] [doi: 10.1197/jamia.M3144] [Medline: 19567795]
8. Sengupta S, Calman NS, Hripcsak G. A model for expanded public health reporting in the context of HIPAA. J Am Med Inform Assoc 2008;15(5):569-574 [FREE Full text] [doi: 10.1197/jamia.M2207] [Medline: 18579843]
9. Willison DJ, Emerson C, Szala-Meneok KV, Gibson E, Schwartz L, Weisbaum KM, et al. Access to medical records for research purposes: Varying perceptions across research ethics boards. J Med Ethics 2008 Apr;34(4):308-314. [doi: 10.1136/jme.2006.020032] [Medline: 18375687]
10. Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, et al. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med Res Methodol 2016 Jul 08;16 Suppl 1:77 [FREE Full text] [doi: 10.1186/s12874-016-0169-4] [Medline: 27410040]
11. Bierer BE, Li R, Barnes M, Sim I. A global, neutral platform for sharing trial data. N Engl J Med 2016 Jun 23;374(25):2411-2413. [doi: 10.1056/NEJMp1605348] [Medline: 27168194]
12. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Investig Med 2010 Jan;58(1):11-18 [FREE Full text] [doi: 10.2310/JIM.0b013e3181c9b2ea] [Medline: 20051768]
13. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: A review of recent research. BMC Med Res Methodol 2010 Aug 02;10:70 [FREE Full text] [doi: 10.1186/1471-2288-10-70] [Medline: 20678228]
14. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. J Biomed Inform 2014 Aug;50:4-19 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.002] [Medline: 24936746]
15. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071 [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
16. Fernández-Alemán JL, Señor IC, Lozoya P, Toval A. Security and privacy in electronic health records: A systematic literature review. J Biomed Inform 2013 Jun;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
17. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Inf Syst 2013 Sep;38(6):946-969. [doi: 10.1016/j.is.2012.11.005]
18. Dankar FK, El Emam K. Transactions on Data Privacy. Catalonia, Spain: IIIA-CSIC; 2013 Apr. Practicing differential privacy in health care: A review URL:https://pdfs.semanticscholar.org/65a5/37c9cd327c2925676f59ddffa01cf4afbe51.pdf [accessed 2019-05-23] [WebCite Cache ID 78aeXVavZ]
19. Lasko TA, Vinterbo SA. Spectral anonymization of data. IEEE Trans Knowl Data Eng 2010 Mar 01;22(3):437-446 [FREE Full text] [doi: 10.1109/TKDE.2009.88] [Medline: 21373375]
20. Knoppers BM, Saginur M. The Babel of genetic data terminology. Nat Biotechnol 2005 Aug;23(8):925-927. [doi: 10.1038/nbt0805-925] [Medline: 16082354]
21. Phillips M, Knoppers BM. The discombobulation of de-identification. Nat Biotechnol 2016 Dec 08;34(11):1102-1103. [doi: 10.1038/nbt.3696] [Medline: 27824850]
22. Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information. Washington, DC: US Department of Health and Human Services; 2019. Cases currently under investigation URL:https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf [accessed 2019-01-24] [WebCite Cache ID 75fKWiXfo]
23. Culnane C, Rubinstein B, Teague V. Health Data in an Open World. Melbourne, Australia: The University of Melbourne; 2017 Dec 18. URL:https://arxiv.org/ftp/arxiv/papers/1712/1712.05627.pdf [accessed 2019-05-12] [WebCite Cache ID 78byKIiYt]
24. Tricco AC, Lillie E, Zarin W, O'Brien K, Colquhoun H, Kastner M, et al. A scoping review on the conduct and reporting of scoping reviews. BMC Med Res Methodol 2016 Feb 09;16:15 [FREE Full text] [doi: 10.1186/s12874-016-0116-4] [Medline: 26857112]
25. Levac D, Colquhoun H, O'Brien KK. Scoping studies: Advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
26. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
27. US National Library of Medicine. 2019. Fact Sheet: MEDLINE® Journal Selection URL:https://www.nlm.nih.gov/lstrc/jsel.html [accessed 2018-05-24] [WebCite Cache ID 78byuh1EF]
28. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care 2012 Jul;50 Suppl:S82-S101 [FREE Full text] [doi: 10.1097/MLR.0b013e3182585355] [Medline: 22692265]

29. Kayaalp M. Patient privacy in the era of big data. Balkan Med J 2018 Dec 20;35(1):8-17 [FREE Full text] [doi: 10.4274/balkanmedj.2017.0966] [Medline: 28903886]
30. Baerlocher MO, Newton M, Gautam T, Tomlinson G, Detsky AS. The meaning of author order in medical research. J Investig Med 2007 May;55(4):174-180. [doi: 10.2310/6650.2007.06044] [Medline: 17651671]
31. Choi HJ, Lee MJ, Choi C, Lee J, Shin S, Lyu Y, et al. Establishing the role of honest broker: Bridging the gap between protecting personal health data and clinical research efficiency. PeerJ 2015;3:e1506 [FREE Full text] [doi: 10.7717/peerj.1506] [Medline: 26713253]
32. Ye H, Chen ES. Attribute Utility Motivated k-anonymization of datasets to support the heterogeneous needs of biomedical researchers. AMIA Annu Symp Proc 2011;2011:1573-1582 [FREE Full text] [Medline: 22195223]
33. Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inform 2014 Apr;83(4):303-312. [doi: 10.1016/j.ijmedinf.2013.11.005] [Medline: 24370391]
34. Pantazos K, Lauesen S, Lippert S. Preserving medical correctness, readability and consistency in de-identified health records. Health Informatics J 2017 Dec;23(4):291-303. [doi: 10.1177/1460458216647760] [Medline: 27199298]
35. O'Neill L, Dexter F, Zhang N. The risks to patient privacy from publishing data from clinical anesthesia studies. Anesth Analg 2016 Dec;122(6):2017-2027. [doi: 10.1213/ANE.0000000000001331] [Medline: 27172145]
36. Grouin C, Rosier A, Dameron O, Zweigenbaum P. Testing tactics to localize de-identification. Stud Health Technol Inform 2009;150:735-739. [Medline: 19745408]
37. Townend D. EU laws on privacy in genomic databases and biobanking. J Law Med Ethics 2016 Dec;44(1):128-142. [doi: 10.1177/1073110516644204] [Medline: 27256129]
38. Sariyar M, Schlünder I. Reconsidering anonymization-related concepts and the term "identification" against the backdrop of the European legal framework. Biopreserv Biobank 2016 Oct;14(5):367-374 [FREE Full text] [doi: 10.1089/bio.2015.0100] [Medline: 27104620]
39. Prasser F, Kohlmayer F, Kuhn KA. Efficient and effective pruning strategies for health data de-identification. BMC Med Inform Decis Mak 2016 Apr 30;16:49 [FREE Full text] [doi: 10.1186/s12911-016-0287-2] [Medline: 27130179]
40. Shin S, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A de-identification method for bilingual clinical texts of various note types. J Korean Med Sci 2015 Jan;30(1):7-15 [FREE Full text] [doi: 10.3346/jkms.2015.30.1.7] [Medline: 25552878]
41. Kohlmayer F, Prasser F, Eckert C, Kuhn KA. A flexible approach to distributed data anonymization. J Biomed Inform 2014 Aug;50:62-76 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.002] [Medline: 24333850]
42. Heatherly RD, Loukides G, Denny JC, Haines JL, Roden DM, Malin BA. Enabling genomic-phenomic association discovery without sacrificing anonymity. PLoS One 2013;8(2):e53875 [FREE Full text] [doi: 10.1371/journal.pone.0053875] [Medline: 23405076]
43. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proc Natl Acad Sci U S A 2010 Apr 27;107(17):7898-7903 [FREE Full text] [doi: 10.1073/pnas.0911686107] [Medline: 20385806]
44. Tamersoy A, Loukides G, Nergiz ME, Saygin Y, Malin B. Anonymization of longitudinal electronic medical records. IEEE Trans Inf Technol Biomed 2012 May;16(3):413-423 [FREE Full text] [doi: 10.1109/TITB.2012.2185850] [Medline: 22287248]
45. Malin B, Benitez K, Masys D. Never too old for anonymity: A statistical standard for demographic data sharing via the HIPAA Privacy Rule. J Am Med Inform Assoc 2011;18(1):3-10 [FREE Full text] [doi: 10.1136/jamia.2010.004622] [Medline: 21169618]
46. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc 2010;17(3):322-327 [FREE Full text] [doi: 10.1136/jamia.2009.002725] [Medline: 20442151]
47. Loukides G, Gkoulalas-Divanis A. Utility-aware anonymization of diagnosis codes. IEEE J Biomed Health Inform 2013 Jan;17(1):60-70. [doi: 10.1109/TITB.2012.2212281] [Medline: 22893444]
48. Prasser F, Kohlmayer F, Kuhn KA. The importance of context: Risk-based de-identification of biomedical data. Methods Inf Med 2016 Aug 05;55(4):347-355. [doi: 10.3414/ME16-01-0012] [Medline: 27322502]
49. Heatherly R, Rasmussen LV, Peissig PL, Pacheco JA, Harris P, Denny JC, et al. A multi-institution evaluation of clinical profile anonymization. J Am Med Inform Assoc 2016 Apr;23(e1):e131-e137 [FREE Full text] [doi: 10.1093/jamia/ocv154] [Medline: 26567325]
50. Goodman D, Johnson CO, Bowen D, Smith M, Wenzel L, Edwards K. De-identified genomic data sharing: The research participant perspective. J Community Genet 2017 Jul;8(3):173-181 [FREE Full text] [doi: 10.1007/s12687-017-0300-1] [Medline: 28382417]
51. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform 2018 Dec;22(2):611-622. [doi: 10.1109/JBHI.2017.2676880] [Medline: 28358693]
52. Prasser F, Bild R, Kuhn KA. A generic method for assessing the quality of de-identified health data. Stud Health Technol Inform 2016;228:312-316. [Medline: 27577394]
53. El Emam K, Moher E. Privacy and anonymity challenges when collecting data for public health purposes. J Law Med Ethics 2013 Mar;41 Suppl 1:37-41. [doi: 10.1111/jlme.12036] [Medline: 23590738]

54. Majeed A, Ullah F, Lee S. Vulnerability- and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data. Sensors (Basel) 2017 May 08;17(5):1-23 [FREE Full text] [doi: 10.3390/s17051059] [Medline: 28481298]
55. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. J Am Med Inform Assoc 2015 Sep;22(5):1029-1041 [FREE Full text] [doi: 10.1093/jamia/ocv004] [Medline: 25911674]
56. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. Trials 2010 Jan 29;11:9 [FREE Full text] [doi: 10.1186/1745-6215-11-9] [Medline: 20113465]
57. Lin W, Yang D, Wang J. Privacy preserving data anonymization of spontaneous ADE reporting system dataset. BMC Med Inform Decis Mak 2016 Dec 18;16 Suppl 1:58 [FREE Full text] [doi: 10.1186/s12911-016-0293-4] [Medline: 27454754]
58. Bartholomäus S, Hense HW, Heidinger O. Blinded Anonymization: A method for evaluating cancer prevention programs under restrictive data protection regulations. Stud Health Technol Inform 2015;210:424-428. [Medline: 25991179]
59. Fernandes AC, Cloete D, Broadbent MT, Hayes RD, Chang C, Jackson RG, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med Inform Decis Mak 2013 Jul 11;13:71 [FREE Full text] [doi: 10.1186/1472-6947-13-71] [Medline: 23842533]
60. El Emam K, Dankar FK, Vaillancourt R, Roffey T, Lysyk M. Evaluating the risk of re-identification of patients from hospital prescription records. Can J Hosp Pharm 2009 Jul;62(4):307-319 [FREE Full text] [Medline: 22478909]
61. Meystre S, Heider P, Kim Y, Trice A, Underwood G. Clinical text automatic de-identification to support large scale data reuse and sharing: Pilot results. In: Proceedings of the AMIA 2018 Annual Symposium. 2018 Nov 05 Presented at: AMIA 2018 Annual Symposium; November 3-7, 2018; San Francisco, CA URL:https://symposium2018.zerista.com/event/member/509567
62. Shaw DM. Blinded by the light: Anonymization should be used in peer review to prevent bias, not protect referees. EMBO Rep 2015 Aug;16(8):894-897 [FREE Full text] [doi: 10.15252/embr.201540943] [Medline: 26174615]
63. Sinnott R, Ajayi O, Stell A, Young A. Towards a virtual anonymisation grid for unified access to remote clinical data. Stud Health Technol Inform 2008;138:90-101. [Medline: 18560111]
64. Lu Y, Sinnott RO, Verspoor K. A semantic-based k-anonymity scheme for health record linkage. Stud Health Technol Inform 2017;239:84-90. [Medline: 28756441]
65. Tamersoy A, Loukides G, Denny JC, Malin B. Anonymization of administrative billing codes with repeated diagnoses through censoring. AMIA Annu Symp Proc 2010 Nov 13;2010:782-786 [FREE Full text] [Medline: 21347085]
66. Thorogood A, Zawati MH. International guidelines for privacy in genomic biobanking (or the unexpected virtue of pluralism). J Law Med Ethics 2015;43(4):690-702. [doi: 10.1111/jlme.12312] [Medline: 26711410]
67. Narayanan A, Felten EW. randomwalker.info. 2014 Jul 09. No silver bullet: De-identification still doesn't work URL:http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf [accessed 2019-01-24] [WebCite Cache ID 75fQcTcoA]
68. Foufi V, Gaudet-Blavignac C, Chevrier R, Lovis C. De-identification of medical narrative data. Stud Health Technol Inform 2017;244:23-27. [Medline: 29039370]
69. O'Keefe CM, Westcott M, O'Sullivan M, Ickowicz A, Churches T. Anonymization for outputs of population health and health services research conducted via an online data center. J Am Med Inform Assoc 2017 May 01;24(3):544-549. [doi: 10.1093/jamia/ocw152] [Medline: 28011594]
70. Fullerton SM, Lee SS. Secondary uses and the governance of de-identified data: Lessons from the human genome diversity panel. BMC Med Ethics 2011 Sep 26;12:16 [FREE Full text] [doi: 10.1186/1472-6939-12-16] [Medline: 21943371]
71. Phillips M, Dove ES, Knoppers BM. Criminal prohibition of wrongful re-identification: Legal solution or minefield for big data? J Bioeth Inq 2017 Dec;14(4):527-539 [FREE Full text] [doi: 10.1007/s11673-017-9806-9] [Medline: 28913771]
72. Kohlmayer F, Prasser F, Kuhn KA. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. J Biomed Inform 2015 Dec;58:37-48 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.007] [Medline: 26385376]
73. Arbuckle L, Moher E, Bartlett SJ, Ahmed S, El Emam K. Montreal Accord on Patient-Reported Outcomes (PROs) use series - Paper 9: Anonymization and ethics considerations for capturing and sharing patient-reported outcomes. J Clin Epidemiol 2017 Sep;89:168-172. [doi: 10.1016/j.jclinepi.2017.04.016] [Medline: 28433677]
74. Smith C. Preventing unintended disclosure of personally identifiable data following anonymisation. Stud Health Technol Inform 2017;235:313-317. [Medline: 28423805]
75. Cimino JJ. The false security of blind dates: Chrononymization's lack of impact on data privacy of laboratory data. Appl Clin Inform 2012;3(4):392-403 [FREE Full text] [doi: 10.4338/ACI-2012-07-RA-0028] [Medline: 23646086]
76. South BR, Mowery D, Suo Y, Leng J, Ferrández Ó, Meystre SM, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform 2014 Aug;50:162-172 [FREE Full text] [doi: 10.1016/j.jbi.2014.05.002] [Medline: 24859155]
77. The European Parliament and the Council of the European Union. Official Journal of the European Union. Volume 59. Luxembourg: Publications Office of the European Union; 2016 May 04. Legislation URL:https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=EN [accessed 2019-05-27] [WebCite Cache ID 78gSMqiNn]

Abbreviations

DUA: data use agreement
GDPR: General Data Protection Regulation
GO FAIR: Global Open Findable, Accessible, Interoperable, and Reusable
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
MBA: Master of Business Administration
MD: Doctor of Medicine
MeSH: Medical Subject Headings
NIH: National Institutes of Health
NLP: natural language processing
ORCID: Open Researcher and Contributor ID
PHI: protected health information
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
[ti]: Title
[tiab]: Title/Abstract

Edited by M Focsa; submitted 24.01.19; peer-reviewed by B Knoppers, AJ Greenberg, J Goris; comments to author 13.02.19; revised version received 29.03.19; accepted 26.04.19; published 31.05.19

Please cite as:
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
URL: http://www.jmir.org/2019/5/e13484/
doi: 10.2196/13484
PMID: 31152528