frontiers | Frontiers in Bioinformatics

TYPE Systematic Review
PUBLISHED 22 December 2022
DOI 10.3389/fbinf.2022.984807

OPEN ACCESS

EDITED BY
Tao Zeng,
Guangzhou Laboratory, China

REVIEWED BY
Boyan Xu,
Guangdong University of Technology, China
Sibghat Ullah Bazai,
BUITEMS, Pakistan

*CORRESPONDENCE
Ali Sepas,
✉ sepasali1024@gmail.com

SPECIALTY SECTION
This article was submitted to Integrative Bioinformatics, a section of the journal Frontiers in Bioinformatics

RECEIVED 02 July 2022
ACCEPTED 28 November 2022
PUBLISHED 22 December 2022

CITATION
Sepas A, Bangash AH, Alraoui O, El Emam K and El-Hussuna A (2022), Algorithms to anonymize structured medical and healthcare data: A systematic review.
Front. Bioinform. 2:984807.
doi: 10.3389/fbinf.2022.984807

Algorithms to anonymize structured medical and healthcare data: A systematic review

Ali Sepas[1,2]*, Ali Haider Bangash[1,3], Omar Alraoui[4], Khaled El Emam[5] and Alaa El-Hussuna[1]

[1]Open Source Research Collaboration, Aalborg, Denmark, [2]Department of Materials and Production, Aalborg University, Aalborg, Denmark, [3]STMU Shifa College of Medicine, Islamabad, Pakistan, [4]Department of Health Science and Technology, Aalborg University, Aalborg, Denmark, [5]Canada Research Chair in Medical AI, University of Ottawa, Ottawa, ON, Canada

Introduction: With many anonymization algorithms developed for structured medical health data (SMHD) in the last decade, our systematic review provides a comprehensive bird's eye view of algorithms for SMHD anonymization.

Methods: This systematic review was conducted according to the recommendations in the Cochrane Handbook for Reviews of Interventions and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). Eligible articles from the PubMed, ACM digital library, Medline, IEEE, Embase, Web of Science Collection, Scopus, ProQuest Dissertation, and Theses Global databases were identified through systematic searches. The following parameters were extracted from the eligible studies: author, year of publication, sample size, and relevant algorithms and/or software applied to anonymize SMHD, along with the summary of outcomes.

Results: Among 1,804 initial hits, the present study considered 63 records including research articles, reviews, and books. Seventy five evaluated the anonymization of demographic data, 18 assessed diagnosis codes, and 3 assessed genomic data. One of the most common approaches was k-anonymity, which was utilized mainly for demographic data, often in combination with another algorithm; e.g., l-diversity. No approaches have yet been developed for protection against membership disclosure attacks on diagnosis codes.

Conclusion: This study reviewed and categorized different anonymization approaches for MHD according to the anonymized data types (demographics, diagnosis codes, and genomic data). Further research is needed to develop more efficient algorithms for the anonymization of diagnosis codes and genomic data. The risk of reidentification can be minimized with adequate application of the addressed anonymization approaches.

Systematic Review Registration: [http://www.crd.york.ac.uk/prospero], identifier [CRD42021228200].

KEYWORDS

anonymization, de-identification, medical health data, reidentification, electronic health records

Introduction

Over the past two decades, increasing medical health data (MHD) have been collected for secondary purposes such as medical research. MHD contains information such as patient demographics, diagnostics, medication history, and, in some cases, family history. MHD is normally stored in databases available to medical researchers (Gkoulalas-Divanis and Loukides, 2015). While these databases allow researchers to research epidemiology, novel treatment quality, register-based cohort studies, etc. (Gkoulalas-Divanis and Loukides, 2015), they have also increased the risk of reidentification (RR) attack (El Emam et al., 2011). A systematic review by Khaled El Imam and colleagues revealed that 34% of reidentification attacks on medical data were successful (El Emam et al., 2011). Although this study was limited to datasets with relatively small sample sizes, RR is clearly a potentially significant threat (El Emam et al., 2011). To minimize the risk of reidentification due to systematic cyber assaults on MHD, researchers have developed sophisticated techniques and algorithms to anonymize data such that the data can be used for secondary purposes while simultaneously maintaining patient anonymity (Langarizadeh et al., 2018). If data are anonymized sufficiently in compliance with ethical guidelines, written patient consent is not required to utilize their data for secondary purposes; thus, the risk of bias due to a consensus from a fraction of patients and not the entire patient population, is eliminated (El Emam and Arbuckle, 2014). What makes anonymization quite tedious is the delicate balance required between data utility and privacy (El Emam and Arbuckle, 2014). If the data are anonymized to such an extent that they provide no beneficial information about patients, the data are rendered useless; conversely, if the data utility is high, the risk of reidentification grows substantially (Sánchez et al., 2014). One approach to anonymization is Datafly, which applies information generalization, insertion, substitution, and removal to deidentify data (Sweeney, 1998). Another widely utilized deidentification method is optimal lattice anonymization (OLA), which utilizes the k-anonymity method and primarily deidentifies quasi-identifiers (El Emam et al., 2009). A relatively novel anonymization approach is Utility-Preserving Anonymization for Privacy Preserving Data Publishing (PPDP), which also applies the k-anonymity technique and comprises three parts: a utility-preserving model, counterfeit record insertion, and a catalogue of counterfeit records (Sánchez et al., 2014). Although many methods have been suggested, all have strengths and limitations. Moreover, it is not clear, how these different methods compare and which approaches are most suitable for achieving anonymization for a specific purpose.

Therefore, this systematic review aimed to analyze the strengths and weaknesses regarding the RR and data utility of algorithms and software that anonymize structured MHD. As a secondary goal, this study aimed to provide medical health researchers and personnel an opportunity to find and utilize the most suitable algorithm/software for their specific goal(s), by giving an overview of currently available anonymization approaches for structured MHD.

Methods

This systematic review was conducted according to a pre-defined study protocol. The review was registered in the International Prospective Register for Systematic Reviews (PROSPERO, http://www.crd.york.ac.uk/prospero, reg. no. CRD42021228200) and was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for systematic reviews (Page et al., 2021).

Search strategy

The PubMed, ACM digital library, Medline, IEEE, Embase, Web of Science Collection and Scopus, ProQuest Dissertation, and Theses Global databases were searched systematically. The systematic search terms were discussed with a librarian from the University of Aalborg, Denmark, to ensure that all relevant keywords were included. Additionally, a manual search in the following journals was conducted: *Studies in Health Technology and Informatics*, the *International Journal of e-Healthcare Information Systems*, and the *Journal of Biomedical Informatics* using the search terms "anonymization of medical health data" or "anonymization". Moreover, manual searches in the reference lists of papers on this topic, contact with experts in bioinformatics, and a campaign using the Twitter and LinkedIn accounts of #OpenSourceResearch collaboration (Open Source Research Organisation, 2021) were used to collect data about any other algorithms/software to ensure that the overview of the subject was as complete as possible. The following keywords were utilized in the systematic search:

- Deidentifi* OR Depersonali* OR Anonymi*
  AND
- Medical Health data OR Medical Health records OR Electronic health data OR Electronic medical records OR Digital health records OR Digital medical data,
  AND
- Data utility, Data usefulness

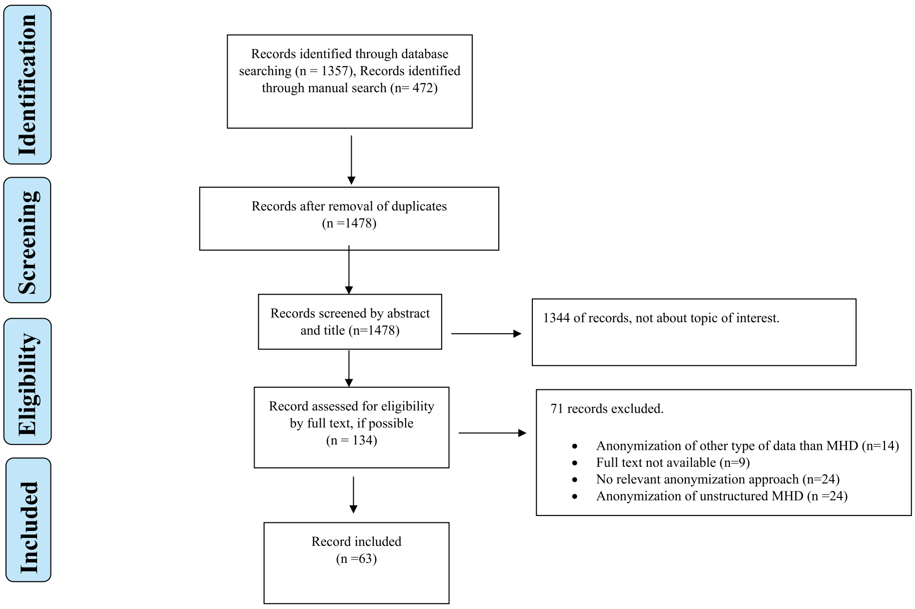
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1357), Records identified through manual search (n= 472)
Records after removal of duplicates (n =1478)
Records screened by abstract and title (n=1478)
1344 of records, not about topic of interest.
Record assessed for eligibility by full text, if possible (n = 134)
71 records excluded.
Anonymization of other type of data than MHD (n=14)
Full text not available (n=9)
No relevant anonymization approach (n=24)
Anonymization of unstructured MHD (n =24)
Record included (n =63)

FIGURE 1
Flowchart.

Inclusion and exclusion criteria

Inclusion criteria

1. Original studies, reviews, and books about anonymization (de-identification) of structured and/or semi-structured medical data for secondary usage.
2. Studies about anonymized medical records that performed assessments of the risk of reidentification and data utility.
3. Studies that applied or investigated de-identification methods and relevant algorithms to anonymize medical data and assessed the risk of reidentification and data utility.

Exclusion criteria

1. To provide the most up-to-date review, studies published before 2000 were excluded. Studies for which the full text was not available were also excluded.
2. Newspaper articles, conference abstracts, and letters to editors were also excluded.

Screening and data extraction

Two researchers (A.S and O.A) independently conducted the screening using the systematic review software Rayyan (Ouzzani et al., 2016). Any disagreements in exclusion or inclusion were resolved by discussion or the involvement of the senior author (A.E). The following parameters were extracted from eligible studies: author, year of publication, sample size, relevant algorithms and computer programs applied to anonymize MHD, and a summary of outcomes. Data extraction was independently conducted by two authors (A.S and O.A. Disagreements were resolved by discussion or the involvement of the senior author (A.E).

Results

The systematic search and manual search identified a total of 1,804 records. Figure 1 shows the PRISMA flowchart. In the initial phase of screening by title and abstract, 1,478 records did not meet the inclusion criteria. Thus, 134 records were assessed for eligibility by full-text screening. A total of 63 records were included in the qualitative analysis (Figure 1), comprising 53 research articles, 8 reviews, and 2 books.

The results suggested that anonymization is most widely applied for protection against identity disclosure, primarily Multi-Sensitive (k, θ*)-anonymity (Liu et al., 2021), with θ* denoting different sensitive values, produced anonymized datasets with low levels of information loss and consistently negligible RR for different values of k and θ* (Liu et al., 2021). Supplementary Table S1 provides a detailed summary of the relevant findings from each record. We divided the different anonymization approaches into three categories (anonymization of demographics, diagnosis codes and genomic data) and three sub-categories based on the attack type that they sought to minimize (identity, membership, and attribute disclosures). Tables 1, 2 provide an overview of the different approaches.

TABLE 1 Algorithms for the anonymization of structured healthcare data pertinent to demographic data.

	Privacy models	Study number
Attack model	Demographics	
Identity disclosure	k-Minimal generalization Gkoulalas-Divanis et al. (2014)	55
	OLA El Emam et al. (2009)	7
	Incognito Gkoulalas-Divanis et al. (2014)	55
	Genetic Gkoulalas-Divanis et al. (2014)	55
	Mondrian Gkoulalas-Divanis et al. (2014)	55
	TDS Gkoulalas-Divanis et al. (2014)	55
	Greedy Gkoulalas-Divanis et al. (2014)	55
	k-member Gkoulalas-Divanis et al. (2014)	55
	KACA Gkoulalas-Divanis et al. (2014)	55
	Agglomerative Gkoulalas-Divanis et al. (2014)	55
	(k,k)-Anonymizer Gkoulalas-Divanis et al. (2014)	55
	Hilb Gkoulalas-Divanis et al. (2014)	55
	iDist Gkoulalas-Divanis et al. (2014)	55
	MDAV Gkoulalas-Divanis et al. (2014)	55
	CBFS Gkoulalas-Divanis et al. (2014)	55
	LSD Mondrian Gkoulalas-Divanis et al. (2014)	55
	NNG Gkoulalas-Divanis et al. (2014)	55
	r-Gather Aggarwal et al. (2010)	1
	Reliability enhancing software in ARX Bild et al. (2020)	3
	Anonymization of multiple sensitive attributes Chester et al. (2020)	4
	Chrononymization Cimino (2012)	5
	Greedy grouping algorithm Cormode et al. (2010)	6
	OLA El Emam et al. (2009)	8
	k-anonymity and l-diversity based anonymizer Gardner and Xiong (2008)	12
	3-anonymity level using ARX Gentili et al. (2017)	13
	Rare disease anonymization using HIPAA safe harbor Gow et al. (2020)	14
	Objective based anonymization in according to HIPAA rules Jung et al. (2018)	19
	Globally optimal algorithm, can be combined with k-anonymity, l-diversity, t-closeness, δ-presence, or many other methods Kohlmayer et al. (2014)	25
	Generalization and Suppression in ARX Kohlmayer et al. (2015)	26
	Generalization with prevention of overgeneralization Lee et al. (2017)	27
	Counterfeit insertion Lee et al. (2017)	27
	Multi-Sensitive (k, θ*)-anonymity Lin et al. (2016)	28
	Clustering by greedy algorithm and k-anonymization Loukides and Jianhua (2006)	33
	Anonymization according to Safe harbor policy and GenEth disclosure policy Malin et al. (2011)	35

(Continued on following page)

TABLE 1 (*Continued*) Algorithms for the anonymization of structured healthcare data pertinent to demographic data.

	Privacy models	Study number
	SDC Martínez et al. (2013)	36
	LKC-privac Mohammed et al. (2009)	38
	SRLA Mohapatra and Patra (2019)	39
	Anonymization with strategies like data swapping, value suppression, generalization, micro aggregation, and noise addition Pika et al. (2020)	41
	De-identification shared task guidelines to longitudinal medical records Stubbs and Uzuner (2014)	45
	HIPAA anonymization rules Tucker et al. (2016)	49
	k-anonymity extension by generalization Ye and Chen (2011)	51
	k-anonymity combined with l-diversity Yoo et al. (2012)	52
	Swapping data anonymization method36	15
	k-anonymity combined with generalization followed by suppression Mawji et al. (2022)	37
	l-diversity slicing approach Onesimu et al. (2022)	40
	Sequential noise addition to event dates k-anonymity with local suppression Templ et al. (2022)	48
Membership disclosure		
	SPALM Gkoulalas-Divanis et al. (2014)	55
	MPALM Gkoulalas-Divanis et al. (2014)	55
	SFALM Gkoulalas-Divanis et al. (2014)	55
	Globally optimal algorithm, can be combined with k-anonymity, l-diversity, t-closeness, δ-presence, or many other methods Gardner and Xiong (2008)	25
	l-diversity slicing approach Onesimu et al. (2022)	40
Attribute disclosure		
	Incognito with l-diversity Gkoulalas-Divanis et al. (2014)	55
	Incognito with t-closeness Gkoulalas-Divanis et al. (2014)	55
	Incognito with (a,k)-anonymity Gkoulalas-Divanis et al. (2014)	55
	p-Sensitive k-anonymity Gkoulalas-Divanis et al. (2014)	55
	Mondrian with l-diversity Gkoulalas-Divanis et al. (2014)	55
	Mondrian with t-closeness Gkoulalas-Divanis et al. (2014)	55
	Top down Gkoulalas-Divanis et al. (2014)	55
	Greedy algorithm Gkoulalas-Divanis et al. (2014)	55
	Hilb with l-diversity Gkoulalas-Divanis et al. (2014)	55
	iDist with l-diversity Gkoulalas-Divanis et al. (2014)	55
	Anatomize Gkoulalas-Divanis et al. (2014)	55
	Delay free anonymization Kim et al. (2014)	23
	Global generalization, local generalization, and bucketization Kim et al. (2017)	24
	Globally optimal algorithm, can be combined with k-anonymity, l-diversity, t-closeness, δ-presence, or many other methods Kohlmayer et al. (2014)	25
	Multi-Sensitive (k, θ*)-anonymity Lin et al. (2016)	28

(Continued on following page)

TABLE 1 (*Continued*) **Algorithms for the anonymization of structured healthcare data pertinent to demographic data.**

	Privacy models	Study number
	HIPAA safe harbor for same disease data (generalization operation utilized) Lin et al. (2016)	29
	LKC-privacy Mohammed et al. (2009)	38
	Closed l-diversification Hsiao et al. (2019)	17
	k-anonymity and l-diversity based anonymizer Gardner and Xiong (2008)	12
	Pseudonymization Somolinos et al. (2015)	44
	k-anonymity combined with l-diversity Yoo et al. (2012)	52
	Combining k-anonymity, l-diversity, and t-closeness Aminifar et al. (2021)	2
	Constraint-based k-means clustering Liu et al. (2021)	30
	l-diversity slicing approach Onesimu et al. (2022)	40

Anonymization of demographic data

A total of 75 algorithms/software were found for the anonymization of demographic data. Some of these approaches were studied in detail by Gkoulalas-Divanis et al. (2014), a brief summary of which is shown in Supplementary Table S1. Forty six approaches were developed for protection against identity disclosure, 5 against membership disclosure, and 24 against attribute disclosure.

Methods against identity disclosure

Identity disclosure is the linkage of an individual or a group of individuals to an entry or a few entries in the dataset. This allows the attacker to obtain highly sensitive data about the exposed individuals. Some of the main approaches are micro aggregation (Domingo-Ferrer and Mateo-Sanz, 2002), generalization (Samarati, 2001), and suppression (Samarati, 2001); however, new approaches such as chrononymization (Cimino, 2012) have also been incorporated. Many of the approaches utilize k-anonymity in combination with other methods to improve the performance, such as Multi-Sensitive (k, θ*)-anonymity (Lin et al., 2016), clustering by greedy algorithm and k-anonymization (Loukides and Jianhua, 2006), and k-anonymity combined with l-diversity (Yoo et al., 2012). Similarly, a delicate balance between privacy protection and data utility was achieved by combining clustering by greedy algorithm and k-anonymization (Loukides and Jianhua, 2006). This algorithm provided better overall data utility than Mondrian; however, the data protection provided by Mondrian was better (Loukides and Jianhua, 2006). The combination of l-diversity and k-anonymity reduced information loss compared to l-diversity and conditional entropy (Yoo et al., 2012). k-anonymity has also been extended by generalization (Ye and Chen, 2011) which showed overall better performance than incognito and Mondrian in terms of lower data distortion with increasing k values, smaller information loss, and a linear decrease of information loss with increasing k[34]. Another approach to counter the issue of overgeneralization is the h-ceiling, in combination with k-anonymity, this method showed a significant reduction in information loss compared to k-anonymity alone. Furthermore, the reconstruction error (RE) was also reduced, the lowest level of information loss was achieved with $h = 0.25$ and the smallest RE with $h = 0.35$. Thus, overall, it was possible to prevent overgeneralization (Lee et al., 2017). Suppression was also applied in ARX software, which showed the lowest level of increase in data utility with a suppression limit of 5%; however, different utility metrics yielded different results (Cimino, 2012). Chrononymization of a single test result could hinder the risk of reidentification, but overall, this approach did not provide sufficient protection against RR (Cimino, 2012).

Methods against membership disclosure

Membership disclosure allows an attacker to determine whether data about a particular individual is contained in a dataset. Protection against this type of attack is more challenging than identity disclosure; consequently, only a handful of approaches have been developed to protect against this type of attack, including SPALM, MPALM, SFALM (Gkoulalas-Divanis et al., 2014), and a globally optimal approach that can be combined with l-diversity, t-closeness, δ-presence, or other methods (Mohammed et al., 2009). Most of the existing algorithms share commonalities with those designed for protection against identity disclosure, such as quasi-identifier transformation and heuristic strategies (Gkoulalas-Divanis et al., 2014). SPALM and MPALM transform quasi-identifiers by generalization and attempt to satisfy δ-presence, while simultaneously minimizing information loss (Nergiz et al., 2007). SPALM generalizes all quasi-identifiers of a similar type in the same way, such as generalizing English as an ethnicity to British. MPALM

TABLE 2 Algorithms for the anonymization of structured healthcare data pertinent to diagnosis codes and genomic data.

	Privacy models	Study number		Study number
Attack model	Diagnosis codes		Genomics data	
Identity disclosure	Combinations Suppression Algorithm Aggarwal et al. (2010)	20	CBA Loukides et al. (2010a)	32
	Clustering based anonymizer (CBA) Bild et al. (2020)	32	ϵ-differentially private mechanism by adopting the *Laplace mechanism* Yu and Ji (2014)	53
	UGACLIP Gkoulalas-Divanis et al. (2014)	55	ϵ-differentially private mechanism by adopting the *exponential mechanism* Yu and Ji (2014)	53
	CBA Gkoulalas-Divanis et al. (2014)	55		
	UAR Gkoulalas-Divanis et al. (2014)	55		
	Apriori Gkoulalas-Divanis et al. (2014)	55		
	LRA Gkoulalas-Divanis et al. (2014)	55		
	VPA Gkoulalas-Divanis et al. (2014)	55		
	mHgHs Gkoulalas-Divanis et al. (2014)	55		
	Recursive partition Gkoulalas-Divanis et al. (2014)	55		
	k-means clustering Lin et al. (2016)	11		
	Anonymization by "dissociation" with application of -anonymity (Loukides et al., 2014)	34		
Attribute disclosure				
	Greedy Gkoulalas-Divanis et al. (2014)	55		
	Suppress control Gkoulalas-Divanis et al. (2014)	55		
	TDControl Gkoulalas-Divanis et al. (2014)	55		
	RBAT Gkoulalas-Divanis et al. (2014)	55		
	Tree-based Gkoulalas-Divanis et al. (2014)	55		
	Sample-based Gkoulalas-Divanis et al. (2014)	55		

generalizes based on context; for instance, English to British in one context and to European in another (Nergiz et al., 2007). SFALM is similar to the previously mentioned approaches but applies c-confident δ-presence; since this approach does not require complete information about the population, it has higher applicability compared to other approaches (Nergiz and Clifton, 2010). The globally optimal approach produced anonymized distributed datasets with information loss ranging between 13% and 87% (Kohlmayer et al., 2014). This approach showed better performance and lower information loss compared to k-anonymity and l-diversity (Kohlmayer et al., 2014).

Methods against attribute disclosure

This type of attack attempts to link individuals to a particular entry (entries) in a data set. One of the most popular methods of protecting against attribute disclosure is l-diversity. Several approaches have been combined with l-diversity, including combination with k-anonymity combined (Yoo et al., 2012), Incognito (Gkoulalas-Divanis et al., 2014), and Hilb (Gkoulalas-Divanis et al., 2014). The combination of k-anonymity and l-diversity provides anonymized datasets with minimum information loss, and less information loss compared to l-diversity alone. Only t-closeness had less information loss than the proposed method; this approach was slower than Entropy l-diversity and t-closeness. l-diversity combined with Incognito also provided anonymization with sufficient utility (Machanavajjhala, 2006; Gkoulalas-Divanis et al., 2014) Hilb with l-diversity (Gkoulalas-Divanis et al., 2014) showed better performance in terms of execution time and information loss compared to Incognito combined with l-diversity (Ghinita et al., 2007a). The approach had lower information loss than Mondrian but had a slower performance (Ghinita et al., 2007b; Gkoulalas-Divanis et al., 2014). Incognito has also been combined with t-closeness (Loukides et al., 2010b). The t-closeness approach attempted to overcome the limitations of l-diversity by requiring that the distribution of an attribute in any equivalence class be close to the distribution of the attribute in the overall table (Ghinita et al., 2007a). t-closeness separated the information gained by an observer from a released table into two parts related to all populations in the data and specific individuals, with the gain of the second type of information gain limited in this approach (Ghinita et al., 2007a). Among other approaches, including global generalization, local generalization, and bucketization (Ye and Chen, 2011), the highest information loss was observed for global generalization, followed by local generalization, and bucketization, where information loss was negligible (Kim et al., 2017). The best overall performance was achieved by Bucketization (Kim et al., 2017). LKC-privacy was developed for larger datasets and was more suitable for blood transfusion service (BTS) data. LKC-privacy allows data sharing, thus providing higher flexibility for BTS data (Mohammed et al., 2009) and higher overall quality of data than k-anonymity (Yoo et al., 2012). For faster anonymization, delay-free anonymization (DF) was developed, which anonymized a single tuple in 0.037 ms compared to 0.18 ms for the accumulated-based method (ABM-1). Information loss by DF was significantly lower than ABM-1, and the l-diverse data set was preserved with a probability of $1/l^{40}$. Pseudonymization is also a novel approach that allows researchers to adjust the relevant parameters for optimal results (Somolinos et al., 2015).

Anonymization of diagnosis codes

The comprehensive systemic search for models of diagnosis code privacy yielded 18 algorithms that aimed to secure diagnosis codes from privacy breaches, unintentional or otherwise. All of these algorithms were related only to identity disclosure. El Emam et al. proposed their Combinations Suppression Algorithm for cases with overlapping combinations of quasi-identifiers and reported less information loss compared to the complete suppression algorithm (Emam et al., 2011). The clustering-based anonymizer (CBA) was presented by Loukides et al. (2010b; Loukides et al. (2010a) for the anonymization of diagnosis codes by clustering and subsequently compared its performance to that of UGACLIP. Comparatively higher satisfaction of utility constraints was reported for CBA with lesser information loss, for the Normalized Certainty Penalty and Average Relative Error (Loukides et al., 2010b). The review by Gkoulalas-Divanis et al. (2014) provided a snapshot of contemporary diagnosis codes privacy algorithms and outlined several key algorithms including, among others, recursive partition, local recoding generalization, and mHgHs. K-means form the basis of a couple of pertinent algorithms related to clustering (Gal et al., 2014) and dissociation anonymization (Gkoulalas-Divanis et al., 2014).

Anonymization of genomic data

A comprehensive search yielded only three privacy algorithms and explored their applications vis-a-vis genomic data based only on the requirements of identity disclosure. The CBA algorithm not only preserved the genomic information but also exhibited superior anonymization capabilities (Loukides et al., 2010b). Yu and Ji (2014) developed algorithms that respectively extended the Laplace and exponential mechanisms and evaluated c^2 statistics and Hamming distance scores to consider the algorithmic performance when applied to a set of single-nucleotide polymorphisms (Yu and Ji, 2014). The superiority of the ∊-differentially private mechanism as extended from the exponential mechanism was demonstrated using the Hamming distance as the score function. However, limitations were demonstrated for the Hamming distance, specifically the early plateau of genomic data utility and the effects of the threshold *p*-value on the data utility (Yu and Ji, 2014).

Discussion

The results of this systematic review demonstrated the feasibility of the anonymization of different types of data such as demographics, diagnosis codes, and genomic data with sufficient levels of protection and utility. The main findings were that for the anonymization of demographics, the combination of classical approaches such as Multi-Sensitive (k, θ*)-anonymity (Lin et al., 2016), extension of k-anonymity with generalization, the combination of k-anonymity with l-diversity (Yoo et al., 2012), and Incognito with l-diversity (Gkoulalas-Divanis et al., 2014) generally provided better data utility and protection than either of methods alone. Issues such as overgeneralization and slow performance were also addressed (Kim et al., 2014; Lee et al., 2017). Moreover, a comparison of some of the algorithms provided researchers an opportunity to select the most suitable anonymization approach for their specific purposes. The findings of this systematic review are consistent with those reported by Langarizadeh et al. (2018) and El Emam and Arbuckle (2014), who concluded that the currently available anonymization approaches provide a delicate balance between data utility and RR for demographic data; however, it is impossible to eliminate RR. Many of the methods are computationally costly, especially for large amounts of data. Pseudonymization was easier to implement for larger data sets and allowed the linkage of data without retaining all identifying characteristics, in contrast to other state-of-the-art approaches.

The potential psychological, financial, and even physical harm to which a patient can be exposed secondary to privacy breaches of diagnosis codes cannot be overstated. Therefore, the optimization of diagnosis code privacy should have paramount significance as the ultimate endpoint for healthcare privacy projects. This snapshot of the diagnosis code privacy-protecting algorithms attempts to reinforce the established considerations among the global healthcare privacy research community including, but not limited, to a heightened recognition of the unambiguous requirement for the development of approaches to optimize statistical analysis capabilities embedded in the information provided by the diagnosis codes with concurrent enhanced suppression of such codes to make it almost impossible for them to be exploited for malicious intent. This may be attained *via* suppressive algorithms that attempt to attenuate utility constraints to a bare minimum. Gkoulalas-Divanis et al. (2014) described the respective pros and cons of different algorithmic models driven by privacy techniques aimed at anonymizing diagnosis codes. Suppression, when employed simultaneously with generalization, provides higher orders of privacy and statistical capabilities compared to those for suppression alone (Gkoulalas-Divanis et al., 2014). Comparisons of bottom-up and top-down heuristic partitioning strategies have demonstrated higher statistical capabilities provided for bottom-up approaches whereas clustering strategies, as those employed by algorithms such as CBA, provide even higher statistical capabilities, although they are computationally expensive (Gkoulalas-Divanis et al., 2014). A strategy to concatenate bottom-up and top-down partitioning strategies has also been reported to optimally provide holistic privacy requirements and concurrently show the superior statistical capabilities provided by diagnosis codes (Gkoulalas-Divanis et al., 2014).

Limitations and directions for future research

The review has some limitations. First, the included studies used different metrics for the assessment of data utility and risk of reidentification, making comparisons of the two approaches challenging, particularly when different metrics were applied for performance evaluation. A wide variety of methods exist for protection against attribute disclosure and identity disclosure, in contrast to the handful of available approaches for protection against membership disclosure. Future research must address this issue, with greater emphasis on protection against membership disclosure. Although anonymization did not provide any apparent advantages over traditional methods (Cimino, 2012), additional research is required to support these findings and further elaborate on the advantages and shortcomings of anonymization. Similarly, pseudonymization is a relatively novel and unexplored domain that requires further investigation, since some clear benefits of this method have been demonstrated (Tinabo et al., 2009). The present study mainly focused on structured MHD; however, novel methods have been developed to handle medical journals and medical images. Our future work aims to also systematically review these anonymization approaches.

Conclusion

In summary, this study reviewed different anonymization approaches for MHD and categorized them according to the anonymized data type (demographics, diagnosis codes, and genomic data). The strengths and limitations of algorithms that protect against identity, attribute, and membership disclosure were addressed. Further research is needed to develop more efficient algorithms for the anonymization of diagnosis codes, and genomic data. The less explored approaches such as chrononymization and pseudonymization yielded promising results of interest for further research. The risk of reidentification can be minimized with adequate application of the included anonymization approaches.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

Study conception: AE and AS. All authors contributed to data collection, analysis, writing, and review of the manuscript.

Acknowledgments

We acknowledge the librarian at Aalborg University, Denmark, for assisting with the creation of the relevant keywords for the systematic search and the modification of these keywords as necessary for different databases. We acknowledge Vedat Atici and Omar Fuad Alraoui for contributing to the systematic search and record screening.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2022.984807/full#supplementary-material

References

Aggarwal, G., Panigrahy, R., Feder, T., Thomas, D., Kenthapadi, K., Khuller, S., et al. (2010). Achieving anonymity via clustering. *ACM Trans. Algorithms* 6, 1–19. doi:10.1145/1798596.1798602

Aminifar, A., Rabbi, F., Pun, V. K. I., and Lamo, Y. (2021). "Diversity-aware anonymization for structured health data," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. doi:10.1109/EMBC46164.2021.9629918

Bild, R., Kuhn, K. A., and Prasser, F. (2020). Better safe than sorry - implementing reliable health data anonymization. *Stud. Health Technol. Inf.* 270, 68–72. doi:10.3233/SHTI200124

Chester, A., Koh, Y. S., Wicker, J., Sun, Q., and Lee, J. (2020). "Balancing utility and fairness against privacy in medical data," in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 1226–1233. doi:10.1109/SSCI47803.2020.9308226

Cimino, J. J. (2012). The false security of blind dates: Chrononymization's lack of impact on data privacy of laboratory data. *Appl. Clin. Inf.* 3, 392–403. doi:10.4338/aci-2012-07-ra-0028

Cormode, G., Srivastava, D., Li, N., and Li, T. (2010). Minimizing minimality and maximizing utility: Analyzing methodbased attacks on anonymized data. *Proc. VLDB Endow.* 3, 1045–1056. doi:10.14778/1920841.1920972

Dankar, F. K., El Emam, K., Neisa, A., and Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC Med. Inf. Decis. Mak.* 12, 66. doi:10.1186/1472-6947-12-66

Davis, J. S., and Osoba, O. (2019). Improving privacy preservation policy in the modern information age. *Health Technol. Berl.* 9, 65–75. doi:10.1007/s12553-018-0250-6

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.* 14, 189–201. doi:10.1109/69.979982

El Emam, K., and Arbuckle, L. (2014). *Anonymizing health data: Case studies and methods to get you started.*

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., et al. (2009). A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inf. Assoc.* 16, 670–682. doi:10.1197/jamia.m3144

El Emam, K., Jonker, E., Arbuckle, L., and Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLoS One* 6, e28071. doi:10.1371/journal.pone.0028071

Emam, K., Paton, D., Dankar, F., and Koru, G. (2011). De-identifying a public use microdata file from the Canadian national discharge abstract database. *BMC Med. Inf. Decis. Mak.* 11, 53. doi:10.1186/1472-6947-11-53

Gadad, V., Sowmyarani, C. N., and Kumar, P. R. (2021). "An effective algorithm for multiple sensitive attributes to preserve data privacy," in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 1 378–383.

Gal, T. S., Tucker, T. C., Gangopadhyay, A., and Chen, Z. (2014). A data recipient centered de-identification method to retain statistical attributes. *J. Biomed. Inf. X.* 50, 32–45. doi:10.1016/j.jbi.2014.01.001

Gardner, J., and Xiong, L. H. I. D. E. (2008). Hide: An integrated system for health information DE-identification. *Proc. - IEEE Symp. Comput. Med. Syst.*, 254–259. doi:10.1109/CBMS.2008.129

Gentili, M., Hajian, S., and Castillo, C. (2017). "A case study of anonymization of medical surveys," in ACM Int. Conf. Proceeding Ser. Part, 77–81.

Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. (2007). "Fast data anonymization with low information loss," in 33rd Int. Conf. Very Large Data Bases, VLDB 2007 - Conf. Proc., 758–769.

Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. (2007). *Fast data anonymization with low information loss.*

Gkoulalas-Divanis, A., Loukides, G., and Sun, G. (2014). Publishing data from electronic health records while preserving privacy: A survey of algorithms. *J. Biomed. Inf.* 50, 4–19. doi:10.1016/j.jbi.2014.06.002

Gkoulalas-Divanis, A., and Loukides, G. (2015). Medical data privacy handbook. *Med. Data Priv. Handb.* doi:10.1007/978-3-319-23633-9

Gow, J., Moffatt, C., and Blackport, J. (2020). Participation in patient support forums may put rare disease patient data at risk of re-identification. *Orphanet J. Rare Dis.* 15, 1–12. doi:10.1186/s13023-020-01497-3

Gunawan, D., Nugroho, Y. S., Maryam, M., and Irsyadi, F. Y. Al. (2021). "Anonymizing prescription data against individual privacy breach in healthcare database," in 2021 9th International Conference on Information and Communication Technology (ICoICT), 138–143. doi:10.1109/ICoICT52021.2021.9527430

Heatherly, R., Rasmussen, L. V., Peissig, P. L., Pacheco, J. A., Harris, P., Denny, J. C., et al. (2016). A multi-institution evaluation of clinical profile anonymization. *J. Am. Med. Inf. Assoc.* 23, e131–e137. doi:10.1093/jamia/ocv154

Hsiao, M. H., Lin, W. Y., Hsu, K. Y., and Shen, Z. X. (2019). "On anonymizing medical microdata with large-scale missing values -A case study with the FAERS dataset," in Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, 6505–6508. doi:10.1109/EMBC.2019.8857025

Jung, J., Park, P., Lee, J., Lee, H., Lee, G., and Cha, G. (2018). *A determination scheme for quasi-identifiers using uniqueness and influence for de-identification of clinical data.* doi:10.1166/jmihi.2020.2966

Kanwal, T., Anjum, A., Malik, S. U., Sajjad, H., Khan, A., Manzoor, U., et al. (2021). A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes. *Comput. Secur.* 105, 102224. doi:10.1016/j.cose.2021.102224

Khan, M. S., Anjum, A., Saba, T., Rehman, A., and Tariq, U. (2021). Improved generalization for secure personal data publishing using deviation. *IT Prof.* 23, 75–80. doi:10.1109/mitp.2020.3030323

Khokhar, R. H., Chen, R., Fung, B. C. M., and Lui, S. M. (2014). Quantifying the costs and benefits of privacy-preserving health data publishing. *J. Biomed. Inf. X.* 50, 107–121. doi:10.1016/j.jbi.2014.04.012

Kim, S., Lee, H., and Chung, Y. D. (2017). Privacy-preserving data cube for electronic medical records: An experimental evaluation. *Int. J. Med. Inf.* 97, 33–42. doi:10.1016/j.ijmedinf.2016.09.008

Kim, S., Sung, M. K., and Chung, Y. D. (2014). A framework to preserve the privacy of electronic health data streams. *J. Biomed. Inf. X.* 50, 95–106. doi:10.1016/j.jbi.2014.03.015

Kohlmayer, F., Prasser, F., Eckert, C., and Kuhn, K. A. (2014). A flexible approach to distributed data anonymization. *J. Biomed. Inf. X.* 50, 62–76. doi:10.1016/j.jbi.2013.12.002

Kohlmayer, F., Prasser, F., and Kuhn, K. A. (2015). The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. *J. Biomed. Inf. X.* 58, 37–48. doi:10.1016/j.jbi.2015.09.007

Kolasa, K., Mazzi, F., Leszczuk-Czubkowska, E., Zrubka, Z., and Pentek, M. (2021). State of the art in adoption of contact tracing apps and recommendations regarding privacy protection and public health: Systematic review. *JMIR mHealth uHealth* 9, e23250. doi:10.2196/23250

Langarizadeh, M., Orooji, A., and Sheikhtaheri, A. (2018). Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review. *Stud. Health Technol. Inf.* 248, 80–87.

Lee, H., Kim, S., Kim, J. W., and Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *BMC Med. Inf. Decis. Mak.* 17, 104–112. doi:10.1186/s12911-017-0499-0

Lin, W. Y., Yang, D. C., and Wang, J. T. (2016). Privacy preserving data anonymization of spontaneous ADE reporting system dataset. *BMC Med. Inf. Decis. Mak.* 16, 58. doi:10.1186/s12911-016-0293-4

Liu, X., Li, X. B., Motiwalla, L., Li, W., Zheng, H., and Franklin, P. D. (2016). Preserving patient privacy when sharing same-disease data. *J. Data Inf. Qual.* 7, 1–14. doi:10.1145/2956554

Liu, Y., Conway, D., Wan, Z., Kantarcioglu, M., Vorobeychik, Y., and Malin, B. A. (2021). De-identifying socioeconomic data at the census tract level for medical research through constraint-based clustering. *AMIA Annu. Symp. Proc.* 2021, 793–802.

Loukides, G., Denny, J. C., and Malin, B. (2010). The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Med. Inf. Assoc.* 17, 322–327. doi:10.1136/jamia.2009.002725

Loukides, G., Gkoulalas-Divanis, A., and Malin, B. (2010). Privacy-preserving publication of diagnosis codes for effective biomedical analysis. *ITAB Corfu Greece* 23, 1–6. doi:10.1109/ITAB.2010.5687720

Loukides, G., and Jianhua, S. (2006). "Towards balancing data usefulness and privacy protection in k-anonymisation," in Proc. - Sixth IEEE Int. Conf. Comput. Inf. Technol. CIT, 2006. doi:10.1109/CIT.2006.184

Loukides, G., Liagouris, J., Gkoulalas-Divanis, A., and Terrovitis, M. (2014). Disassociation for electronic health record privacy. *J. Biomed. Inf. X.* 50, 46–61. doi:10.1016/j.jbi.2014.05.009

Machanavajjhala, A. (2006). *Diversity : Privacy beyond k -anonymity.*

Malin, B., Benitez, K., and Masys, D. (2011). Never too old for anonymity: A statistical standard for demographic data sharing via the hipaa privacy rule. *J. Am. Med. Inf. Assoc.* 18, 3–10. doi:10.1136/jamia.2010.004622

Martínez, S., Sánchez, D., and Valls, A. (2013). A semantic framework to protect the privacy of electronic health records with non-numerical attributes. *J. Biomed. Inf. X.* 46, 294–303. doi:10.1016/j.jbi.2012.11.005

Mawji, A., Longstaff, H., Trawin, J., Dunsmuir, D., Komugisha, C., Novakowski, S. K., et al. (2022). A proposed de-identification framework for a cohort of children presenting at a health facility in Uganda. *PLOS Digit. Health* 1, e0000027. doi:10.1371/journal.pdig.0000027

Mohammed, N., Fung, B. C. M., Hung, P. C. K., and Lee, C. K. (2009). "Anonymizing healthcare data: A case study on the blood transfusion service," in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 1285–1293. doi:10.1145/1557019.1557157

Mohapatra, D., and Patra, M. R. (2019). "A graph based approach for privacy preservation of citizen data in e-governance applications," in ACM Int. Conf. Proceeding Ser, 433–438. doi:10.1145/3325112.3325254

Nergiz, M. E., Atzori, M., and Clifton, C. (2007). "Hiding the presence of individuals from shared databases," in Proc. ACM SIGMOD Int. Conf. Manag. Data, 665–676. doi:10.1145/1247480.1247554

Nergiz, M. E., and Clifton, C. (2010). Presence without complete world knowledge. *IEEE Trans. Knowl. Data Eng.* 22, 868–883. doi:10.1109/tkde.2009.125

Olatunji, I. E., Rauch, J., Katzensteiner, M., and Khosla, M. (2021). *A review of anonymization for healthcare data. Big data.* doi:10.1089/big.2021.0169

Onesimu, J. A., K., J., Eunice, J., Pomplun, M., and Dang, H. (2022). Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access* 10, 86979–86997. doi:10.1109/access.2022.3199433

Open Source Research Organisation (2021). Implementing information technologies in medical research. Available at https://osrc.network/.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Syst. Rev.* 5, 210. doi:10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., and Mulrow, C. D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 371. doi:10.1136/bmj.n71

Pika, A., Wynn, M. T., Budiono, S., ter Hofstede, A. H., van der Aalst, W. M., and Reijers, H. A. (2020). Privacy-preserving process mining in healthcare. *Int. J. Environ. Res. Public Health* 17, 1612. doi:10.3390/ijerph17051612

Poulis, G., Loukides, G., Skiadopoulos, S., and Gkoulalas-Divanis, A. (2017). Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints. *J. Biomed. Inf. X.* 65, 76–96. doi:10.1016/j.jbi.2016.11.001

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* 13, 1010–1027. doi:10.1109/69.971193

Sánchez, D., Batet, M., and Viejo, A. (2014). Utility-preserving privacy protection of textual healthcare documents. *J. Biomed. Inf. X.* 52, 189–198. doi:10.1016/j.jbi.2014.06.008

Somolinos, R., Munoz, A., Hernando, M. E., Pascual, M., Caceres, J., Sanchez-de-Madariaga, R., et al. (2015). Service for the pseudonymization of electronic healthcare records based on ISO/EN 13606 for the secondary use of information. *IEEE J. Biomed. Health Inf.* 19, 1937–1944. doi:10.1109/jbhi.2014.2360546

Stubbs, A., and Uzuner, O. (2014). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *J. Biomed. Inf. X.* 58, 20–29. doi:10.1016/j.jbi.2015.07.020

Sweeney, L. (1998). *Datafly: A system for providing anonymity in medical data*, 356–381. doi:10.1007/978-0-387-35285-5_22

Tamersoy, A., Loukides, G., Nergiz, M. E., Saygin, Y., and Malin, B. (2012). Anonymization of longitudinal electronic medical records. *IEEE Trans. Inf. Technol. Biomed.* 16, 413–423. doi:10.1109/titb.2012.2185850

Templ, M., Kanjala, C., and Siems, I. (2022). Privacy of study participants in open-access health and demographic surveillance system data: Requirements analysis for data anonymization. *JMIR Public Health Surveill.* 8, e34472. doi:10.2196/34472

Tinabo, R., Mtenzi, F., and O'Shea, B. (2009). *Anonymisation Vs. Pseudonymisation: Which one is most useful for both privacy protection and usefulness of e-healthcare data.* New York City: ICITST.

Tucker, K., Branson, J., Dilleen, M., Hollis, S., Loughlin, P., Nixon, M. J., et al. (2016). Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med. Res. Methodol.* 16, 77. doi:10.1186/s12874-016-0169-4

Wu, L., He, H., and Zaïane, O. R. (2013). "Utility of privacy preservation for health data publishing," in Proc. CBMS 2013 - 26th IEEE Int. Symp. Comput. Med. Syst., 510–511. doi:10.1109/CBMS.2013.6627853

Ye, H., and Chen, E. S. (2011). "Attribute Utility Motivated k-anonymization of datasets to support the heterogeneous needs of biomedical researchers," in AMIA Annu. Symp. Proc., 1573–1582.

Yoo, S., Shin, M., and Lee, D. (2012). An approach to reducing information loss and achieving diversity of sensitive attributes in k-anonymity methods. *Interact. J. Med. Res.* 1, e14. doi:10.2196/ijmr.2140

Yu, F., and Ji, Z. (2014). Scalable privacy-preserving data sharing methodology for genome-wide association studies: An application to iDASH healthcare privacy protection challenge. *BMC Med. Inf. Decis. Mak.* 14, S3. doi:10.1186/1472-6947-14-s1-s3

Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C., and Al Moubayed, N. (2021). Data anonymization for pervasive health care: Systematic literature mapping study. *JMIR Med. Inf.* 9, e29871. doi:10.2196/29871