Protecting Anonymity in Data-Driven Biomedical Science

Peter Kieseberg[1,2], Heidelinde Hobel[1], Sebastian Schrittwieser[3],
Edgar Weippl[1], and Andreas Holzinger[2]

[1] Secure Business Austria Research
[2] Research Unit HCI, Institute for Medical Informatics, Statistics & Documentation, Medical University Graz
[3] University of Applied Sciences St. Pölten
{firstletteroffirstname,lastname}@sba-research.org,
andreas.holzinger@medunigraz.at
sebastian.schrittwieser@fhstp.ac.at

Abstract. With formidable recent improvements in data processing and information retrieval, knowledge discovery/data mining, business intelligence, content analytics and other upcoming empirical approaches have an enormous potential, particularly for the data intensive biomedical sciences. For results derived using empirical methods, the underlying data set should be made available, at least during the review process for the reviewers, to ensure the quality of the research done and to prevent fraud or errors and to enable the replication of studies. However, in particular in the medicine and the life sciences, this leads to a discrepancy, as the disclosure of research data raises considerable privacy concerns, as researchers have of course the full responsibility to protect their (volunteer) subjects, hence must adhere to respective ethical policies. One solution for this problem lies in the protection of sensitive information in medical data sets by applying appropriate anonymization. This paper provides an overview on the most important and well-researched approaches and discusses open research problems in this area, with the goal to act as a starting point for further investigation.

Keywords: Anonymization, pseudonymization, data-driven sciences, big data, privacy, security, safety.

1 Introduction

New and advanced methods in statistical analysis and rapidly emerging technological improvements, e.g., in computation performance, data storage, cloud computing and technologies that support worldwide collaborative work, have laid the foundation for a new field of science that we call *data-driven sciences*. In particular biomedical informatics is becoming such a data-driven science due to the increasing trend toward personalized and precision medicine [1], [2], [3]. A recent example can be found in [4].

A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 301–316, 2014.
© Springer-Verlag Berlin Heidelberg 2014

Data-driven science uses these new resources to analyze enormous data sets, often called *big data* [5], and reasons based on the empirical findings and evidence from these analyses. The sources of big data can extremely vary, ranging from data gathered online from open sources to data sets provided by research partners, companies or volunteers, or coming from the own laboratories or hospitals; in the medical domain data can come from clinical patient treatment and/or from biomedical research, from hospital sources or from biobanks. The size and complexity of the data sets allows a large variety of inferences to be made, which makes big data very useful for research but can, at the same time, potentially be exploited to extract information that could be used in malicious ways or that might infringe on the privacy of the data subjects [6]. This especially concerns data-driven science in the medical sector, since, as a principle, most data in this field is sensitive and issues of privacy, security, safety and data protection are always an issue [7]. Even when access to the data is limited to a specific researcher or a research team whose members might be from different organizations or universities there is a high risk of disclosure. The more people have access to classified information, the higher the risk of it being exploited for malicious purposes.

However, research, particularly non-commercial research, is usually intended - or should be intended - for public dissemination through conferences or journals [8]. The peer-review procedure normally ensures the quality of such research, but without access to the underlying data, work in the field of data-driven medical science cannot be validated by reviewers. The result is an extremely problematic situation where authors either include the data only in a condensed or abstracted form, which protects privacy but means has the drawback the reader cannot validate the results or evaluate the data for a personal learning effect, or publish the data, even if only for the duration of the review process and with restricted access. The former solution is problematic in that the research cannot be properly reviewed, which results in chances for fraud and poor research, especially in the "publish or perish" atmosphere of pressure to publish frequently to gain the recognition of the academic field or funding institutions.

Furthermore, even in the absence of fraud, missing data can make perfectly valid results look debatable. The latter solution, while mitigating these problems, exposes data sets to potential misuse. Furthermore, especially regarding data-driven research in medical sciences, the publication of the raw data will most certainly result in legal issues. Recently, there is a strong movement towards the promotion of open data sets in biomedical research [9], but what to do in case the data *cannot* be made openly available?

We address the question of how to utilize and share research data without exposing it to risks by providing an overview on the latest anonymization and pseudonymization techniques, following previous work [10] and especially considering the biomedical sector. Furthermore, we will give an overview on open questions, again especially targeting this sensitive field of information processing.

2 Glossary and Key Terms

This section shall define the most important or ambiguous terms used in the paper to avoid any danger of misinterpretation and to ensure a common understanding.

Anatomization: An operation for achieving anonymization, this works by splitting the attributes of table records into QIs and sensitive columns which are stored in separate tables. Then the linking between the two tables in made ambiguous for providing anonymity (see Section 5.1).

Anonymization: A term denoting the removal of personal information from data including the ability to link the data to persons by utilizing characteristics.

Big Data: While this term is currently very popular, there exists no exact definition for it. Usually it is used to either describe the processing of large amounts of data, or as a paradigm for data processing, where information from several sources is mashed up in order to generate additional knowledge.

Business Intelligence (BI): This term describes a selection of methodologies, technologies and architectures for harvesting information relevant for business from raw data.

Data Linkage: The effort of constructing relationships between sensitive published data and data that is public or easily accessible for attackers is called *data linkage*.

Data Precision Metric (DPM): Metric or norm for measuring the information loss due to techniques for anonymization. Usually used in the context of *generalization*.

Generalization: This method replaces sensitive values with more general ones by grouping values in an interval or by using taxonomies and replacing the values with parent nodes, thus reducing the granularity of quasi identifiers (see Section 5.1).

Identifier: An attribute that uniquely identifies a person.

***k*-anonymity:** A paradigm for anonymization. A more detailed description is given in Section 5.2.

***l*-diversity:** This is an extension of the *k*-anonymity paradigm incorporating diversity into the equivalence classes (see Section 5.3).

Permutation: A concept similar to anatomization, this operation also splits records into QIs and sensitive attributes, stores them in different tables and makes the linkage ambiguous (see Section 5.1).

Perturbation: Distortion of data using mechanisms like adding noise or the introduction of synthetic values, further details can be found in Section 5.1.

Pseudonymization: Every identifier and all relevant quasi identifiers are exchanged for pseudonyms in order to cover the identity of the persons in questions.

Quasi Identifier: This are attributes which are not directly identifiers, but can be used in combination to identify persons.

Rule Mining: This keyword covers a multitude of techniques for the automatic extraction of rules in (often large) data sets. It constitutes an important set of techniques in the area of machine learning.

Structured Data: In general, a representation of information that is following fixed rules. Usually used for tables or structured file formats like XML.

Suppression: Single data elements, e.g. rows in a table, are removed from the set in order to get a higher level of anonymization. This technique is often used in order to achieve k-anonymity or related concepts, see Section 5.1.

***t*-closeness:** An extension of l-diversity that is secure against skewness attacks, see Section 5.4.

3 Background

In [11–16] the authors claim that data-driven research is a paradigm that is constantly gaining popularity in most research areas. The term "big data" originated in the IT sector, where large data samples had to be analyzed, usually in order to evaluate proposed algorithms or prototypes, especially with regard to practical applicability and performance issues. They can also be analyzed to derive new empirical findings concerning general trends and characteristics. Health data publishing is a prime example of an area where sensitive data must be protected from leaking into the public [17, 18]. This field has shown that not only direct identifiers, such as social security numbers, can contribute to the threat of a privacy breach, but also so-called quasi-identifiers (QI), e.g., the triple ZIP-code, birth date and gender. It was shown in [17, 18] that this data triple alone allows the unambiguous identification of roughly 80% of the American citizens, resulting that private data, such as illnesses or treatment, can be inferred about them and used for malicious purposes. The effort of constructing relationships between sensitive published data and data that is public or easily accessible for attackers is called *data linkage* [19]. This is not only an issue in health care, either. For example, Dey et al. [13] analyzed approx. 1,400,000 Facebook account settings to infer privacy trends for several personal attributes. Although they used public accounts for their research, their results combined with the data they measured and recorded are highly sensitive and should not be published without applying appropriate anonymization or pseudonymization techniques. We, as researchers, are responsible for protecting the data we use and for preserving the privacy of our subjects, who are often volunteers. This protection includes ensuring the unlinkability of sensitive data so that data sets can be published to allow the validation of research, collaboration between several research groups, and learning by enabling the reader to repeat the described data analysis.

4 Privacy Threats in Data-Driven Science

The set of attributes that comprise research data can usually be divided into several categories: Identifiers, quasi identifiers, as well as sensitive and non-sensitive attributes and inferred knowledge obtained during research.

Identifiers are attributes that more or less uniquely identify a person. Typically, names are considered to be identifiers, even though they rarely "uniquely identify" a person in general (many popular names do not even uniquely identify a person in a small city), as well as addresses. While this definition does lack mathematical rigour, in general there is no big dispute on what is considered to be an identifier. Still, especially in medical data-driven research, we do see a problem with this definition when it comes to genome data, which should be classified as an identifier in our opinion. Considering the above-mentioned Facebook example, we assume that each data record comprising all required data of an account has been classified according to predefined privacy categories. This categorizes the links to the accounts into the *identifier*-category, which has to be removed before publishing.

Quasi-Identifiers (QIs) are a category initially proposed by Dalenius in [20] that includes all attributes that either themselves or in combination could also be used to identify persons. While this definition explicitly includes the identifier-category, these attributes are often removed in current literature, reducing this category to the set of attributes that do not uniquely identify a person themselves, but can pose a danger to privacy if combined with other quasi-identifiers. Common examples for QIs include birthdates or ZIP-codes.

Inference Attacks describe attacks, where background knowledge, a-priori-knowledge or public data sets are used to identify data record owners. This type of attacks is also called *linkage attacks.* In the Facebook example, identifying a person behind an account by mapping data derived from the list of friends to other sources, e.g. students-lists of universities, would result in an inference attack. Commonly, linkage attacks are categorized in four different types: Record linkage, attribute linkage, table linkage and probabilistic attacks [19]. In a *record linkage* attack, the QI is linked directly with additional information, as in the Facebook example described above. *Attribute linkage* looks for a correlation of QIs and inferred knowledge. For example, if an attacker knows a given individual is in a certain equivalence group, they can easily identify that persons sensitive attribute. *Table linkage* attacks determine the presence or absence of the record owner, while *probabilistic attacks* refer to the threat of a change in the general probabilistic belief of the attacker after seeing the published data.

In privacy-preserving data publishing, the identity of the record owners is usually hidden using anonymization [21] of the quasi-identifiers to prevent linkage without major information loss. There exist a number of ways in which data can be anonymized. The simplest method is the removal of attributes (quasi-identifiers) before publication in order to increase the difficulty of correctly re-identifying the individual. However, this can prevent the validation of the research method if the removed attributes influence the inferred knowledge. Additionally, the general objective of data-driven science is to find comprehensive knowledge, which means that a change in the "probabilistic belief" of an attacker is unavoidable. It can also be difficult to identify all attributes that constitute QIs, rendering anonymization efforts incomplete and, in the worst case, ineffectual.

Another closely related problem results from the new "Big Data" paradigm, where massive amounts of data from various sources are combined in order to mine correlations and/or derive rules. This is especially sensitive in case of open data initiatives, where data vaults are opened for the public and data from various sources can be combined through mashups, without prior verification of the resulting combination's sensitivity. A more detailed explanation of the inherent problems of this approach, together with a concept solution can be found in [22].

5 Anonymization Concepts

In this chapter we will discuss a selection of operations that can be used for achieving anonymization, followed by a selection of the most popular and well-researched models for defining anonymity.

5.1 Anonymization Operations

Methods of anonymization often relate to the removal or replacement of quasi-identifiers, making the relationship between QIs and sensitive values or inferred knowledge ambiguous, and distorte the data. Fung et al. [19] provided a survey on state-of-the-art anonymization techniques, which can be divided into the following categories: Suppression, generalization, anatomization, permutation and perturbation.

Suppression is the removal or replacement of data tuples, e.g. rows of a table, before publishing. While being the most basic method, it can help yielding good results and is often used together with generalization for achieving k-anonymity. Still, the removal of data rows may lead to a drastic change in the significance of the underlying data, especially when studying rare diseases. Thus, this method must be selected with great care and it must be made sure that the removed values do not change the distribution of relevant attributes significantly. Besides this basic definition of suppression, also called *Record Suppression*, that relies on the removal of whole records, some modified approaches have been proposed: Sometimes it is needed to suppress every appearance of a given value in a table (see [23]) or suppressing single cells (see [24]).

Generalization also replaces values, but seeks to preserve as much information as possible while meeting the requirements of the chosen privacy model. This method replaces sensitive values with more general ones by grouping values in an interval or by using taxonomies and replacing the values with parent nodes, e.g. classifying a date such as 01.01.1999 in the interval $[1990 - 1999]$ or generalizing "Jane Doe" as "female". Figure 1 shows two possible generalization strategies for different kinds of attributes. The actual information loss is measured using so-called *Data Precision Metrics (DMPs)*[1]. While suppression is applied to single

[1] Despite the terminology, most DPMs are not metrics in a mathematical sense.

data elements (i.e. table rows), generalization affects entire attribute classes (i.e. table columns). In general, generalization is the method most often found in the literature on anonymization strategies. Still, there exist several extensions of this concept, e.g. *cell generalization* as introduced by LeFevre et. al. in [25] or *multi-dimensional generalization* (see [26, 27]).

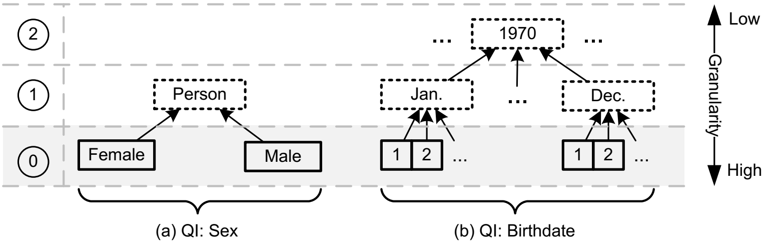

Fig. 1. Generalization of quasi-identifiers

Anatomization makes the relationship between QIs and inferred knowledge ambiguous by grouping, thereby solving the problem illustrated in Table 2. This works by splitting the quasi identifiers and the sensitive attributes into two tables T_Q, holding the quasi-identifiers, and T_S, containing the sensitive values, while adding a shared attribute *id* to both. Now the quasi identifiers are generalized in a way to make the linking between the two tables ambiguous - each characteristic of the sensitive data should then be linkable to each of the l classes of quasi identifiers, where l is a fixed threshold that determines the level of unlinkability. The main advantage of this concept is that the table holding the sensitive values can remain far more detailed compared to a pure generalization based approach, thus making them more valuable for statistic analysis.

Permutation is a concept rather similar to anatomization. It also relies on ambiguity, shuffling inferred knowledge tuples into predefined and generalized groups in a way that does not affect statistical evaluation, so that the results are the same before and after the permutation process.

Perturbation distorts data using different techniques, such as adding noise, swapping values or using synthetic values to replace real ones. Many of these methods can be seen as some kind of dual strategy to suppression. While the latter tries to achieve anonymization by removing records from the data set, many perturbation methods add new records. One advantage of many perturbation methods lies in the preservation of statistical information [19], especially when considering techniques that exchange real data for synthetic values, however, especially when searching for unexpected correlations e.g. by applying rule-mining, perturbation may influence the result.

5.2 k-anonymity

The anonymization concept called k-anonymity with its special consideration of quasi-identifiers was first introduced by Sweeney in [17]. She showed that it was possible to identify individuals even after uniquely identifying attributes such as the name or social security number were removed from health records by linking attributes such as ZIP code, birthdate, and sex.

Table 1. Original data and two anonymized sets ($k = 2$)

Original data				First Set			Second Set		
name	sex	birthdate	disease	sex	birthdate	disease	sex	birthdate	disease
Bill	m	01.05.1972	cancer	M	1972	cancer	P	03.1972	cancer
Dan	m	20.05.1972	cancer	M	1972	cancer	P	03.1972	cancer
Anne	f	10.03.1972	anorexia	F	1972	anorexia	P	04.1972	anorexia
Jill	f	31.03.1972	typhlitis	F	1972	typhlitis	P	04.1972	typhlitis

The criterion of k-anonymity is satisfied if each record is indistinguishable from at least $k-1$ other records with respect to the QIs. This means that quasi-identifying attributes must have the same values within a so-called equivalence class (which contains a minimum of k records), so that it is no longer possible to uniquely link an individual to a specific record in that class. This criterion can, e.g., be achieved by generalizing data of quasi-identifiers, such as generalizing the birthdate attribute by giving only the month and the year, or even just the year or decade. High levels of anonymity are possible with this method by raising the value of the threshold k, but lower anonymity levels are often necessary to preserve the significance of the data.

Today, k-anonymity is a widely adopted anonymization method. Over the past years, several improvements have been proposed that introduce new, stricter criteria for k-anonymity, but do not replace the original idea.

5.3 l-diversity

Table 2 illustrates a major limitation of k-anonymity. In this example ($k = 3$), there are three male patients who were all born in 1972. The original k-anonymity algorithm creates an equivalence class for these three records to fulfill the $k = 3$ criterion, making them indistinguishable from each other with respect to the quasi-identifiers. The problem here, however, is that the sensitive attribute (disease) is identical for all three, which effectively negates the anonymization effort. If the sensitive attributes in an equivalence class lack diversity, k-anonymity cannot ensure privacy. This problem can be countered with l-diversity, which requires each equivalence class to have at least l well-represented values for each sensitive attribute [28].

The definition of ”well-represented” depends on the actual data. There are five different approaches, of which the most basic, ”entropy l-diversity”, requires each

Table 2. Original data and two anonymized sets ($k = 2$)

Original data				k-anonymity		
name	sex	birthdate	disease	sex	birthdate	disease
Bill	m	01.05.1974	cancer	M	1974	cancer
Dan	m	20.05.1974	cancer	M	1974	cancer
Anne	f	10.03.1972	anorexia	F	1972	anorexia
Jill	f	31.03.1972	typhlitis	F	1972	typhlitis
William	m	10.12.1974	cancer	M	1974	cancer
Mary	f	12.12.1973	short breath	F	1972	short breath

equivalence class to include at least l different values for the sensitive attributes. Table 2 shows an example for data obeying entropy-l-diversity. To achieve higher levels of entropy diversity, the quasi-identifiers must be further generalized . The main problem of l-diversity is that it only prevents unique matching of an individual to a sensitive value while ignoring the overall distribution of sensitive values. This makes statistical analysis of equivalence classes possible (skewness attack): Consider a microdata set anonymized with entropy-2-diversity that has an equivalence class containing a sensitive attribute that applies only to a very small percentage (e.g., 1%) of a countrys population, e.g. a rare disease. The probability of a specific individual in this equivalence class suffering from this rare disease is up to 50%, which is much higher than the actual probability within the entire population.

5.4 *t*-closeness

The concept of t-closeness was developed as an improvement to k-anonymity and l-diversity in order to mitigate above mentioned skewness attacks. The basic principle lies in choosing the equivalence classes in a way that the distribution of any sensitive attribute in any class is similar to its distribution in the original table [29]. More precisely, let D_{all} be the distribution of a sensitive attribute in the original table holding all data records and D_i be the distribution of that same attribute in the i^{th} equivalence class, for all classes $i = 1 \ldots n$ as defined in the k-anonymity paradigm. Then these equivalence classes are obeying the t-closeness criteria for a given value t if and only if the the distance between D_{all} and D_i is at most $t, \forall i = 1 \ldots n$. However, the main questions is, how to measure this distance between equivalence classes, while including the semantic distance between values. The solution is the so-called *Earth Mover Distance EMD* as defined in [29].

The t-closeness paradigm has some drawbacks tough: (i) The first and most important drawback considers the impact of enforcing t-closeness on the data set: When assuming t-closeness, the sensitive values will have the same distribution in all equivalence classes with respect to the quasi identifiers, thus having a significant impact on the correlation between these attributes and the QIs. Since a lot of research in medical data-driven science is actually targeting at such correlations, t-closeness remains unusable in these cases. (ii) Another drawback is that t-closeness lacks the ability to specify separate protection levels for each quasi identifier. Furthermore, (iii) there still exist special attribute linkage attacks on t-closeness when utilizing it on sensitive numerical attributes as shown in [30].

5.5 Pseudonymization

Pseudonymization is a method related to anonymization that combines the advantages of anonymization and transparency for the publisher [21]. It is frequently employed in research that uses medical records and has the advantage of making it possible to reverse the anonymization process if necessary, e.g. for health care reasons. For example, in [31], pseudonyms are used to implement the link between individual and sensitive data, in this case medical health records. Cryptographic keys ensure that only authorized persons can re-identify the links. Related approaches can also be found in [32] and [33]. In [34], two solutions for protecting sensitive radiology data through pseudonymization are discussed: In the first approach the unique patient identification numbers are exchanged for reversible pseudonyms by using hashing and encryption techniques, the second one works by applying irreversible one-way pseudonyms. Both solutions lead to pseudonymized health records that can be used for research purposes while ensuring patient privacy.

6 Open Problems and Future Trends

Over the last years, a strong trend towards data-driven research methods has emerged in medical science. Results from the analysis of these data sets are improving constantly, which leads to the conclusion that data-driven research approaches will gain even more attention over the next years. For handling medical data there exist clear regulatory frameworks, e.g. the Health Insurance Portability and Accountability Act (HIPAA), which defines to what extent information can be released to third parties and forbids the disclosure of "individually identifiable health information". As medical data is complex and inhomogeneous, there exist many different potential anonymization approaches, while ongoing research in anonymization and pseudonymization promises even better privacy-preserving methods for the future.

Still, there are many problems to be solved in the realm of data protection in data-driven medical science. To begin with, many of the procedures currently in use only work with structured data, e.g. database records. This specially holds true for all techniques based on the k-anonymity concept, where each information particle falls into a well-defined category (column). But even for this heavily structured data, there exist several open research questions, which we will discuss in this chapter, grouped by their assets.

6.1 Questions Regarding Quasi Identifiers

The first group of concerns lies in the selection and treatment of the quasi identifiers. While this is rather straightforward in the standard examples (e.g. sex, birthdate), there are some inherent questions that need discussion, especially relating to the medical sector:

Definition of quasi identifiers. While sounding trivial and indeed very easy to decide in the standard example, this is not so trivial when considering medical data. A diagnose could, for example, be so rare that the field together with the ZIP-code results in deanonymization of the respective person. The diagnose would need to be treated as a QI in this example. Further examples include rare blood types and parts of genome sequences. The determination of QIs is an important research area for guaranteeing the privacy of patients in data-driven medical research.

Generalization of non-trivial QIs. Following the example in the previous paragraph, the field "diagnose", too, is a quasi identifier that a generalization strategy is needed for. While generalization is rather trivial for standard attributes like dates or numbers, it is rather difficult for free text fields. In the case of a diagnose, ontologies could help generating a generalization tree, still, when considering free text like it is found in notes from GPs, a lot of further research is needed.

Independence of QIs. Sometimes, QIs may not be as independent as they seem. E.g., considering the QI "sex", an entry in the field "diagnose" containing "breast cancer" leads to an approx. 99% chance of this record belonging to a female person, thus rendering the generalization of "sex" practically useless without the generalization of the field "diagnose". The research in this sector also includes the identification of such QIs, preferably without too much knowledge required on the respective field. One fruitful approach could be the utilization of rule mining in order to derive such dependencies.

6.2 Questions Regarding Collaboration

Collaborating with other research institutes again opens up several interesting research questions that need to be tackled in the close future.

Data Precision Metrics. Most metrics for measuring the information loss currently in use are rather trivially depending on the selected generalization strategies. While there exist examples for metrics depending on the actual data distribution, these are rather inefficient during the calculation. Finding efficient and expressive metrics seems to be a valuable research question to us. This also includes the question of fairness when using different strategies for different data recipients on the same source data.

Leak Detection. Even in case of perfect privacy protection, the leaking of research data may result in severe damage to the data owner, e.g. due to premature publication of results and the need for subsequent revision, or simply because of the value of the data set itself. While techniques for watermarking databases (e.g. see [35], [36] can be utilized, these are usually independent from the data (not the data storage, though) itself, thus making them removable without reducing the overall quality of the data sets. Thus, research on how to combine leak detection with privacy protection could enhance the willingness of data owners to share their data with other researchers. While there has been some research regarding this during the last years, these approaches (e.g. [37], [38]) currently only cover the basic k-anonymity concept.

6.3 General Research Questions

This Section contains some other research questions related to the topic of protecting privacy in data-driven medical science, which did not fit into the above categories.

Structuring unstructured data. While a lot of data used in medical research naturally possesses the form of structured data (e.g. derived from machines), there is also a wide variety of unstructured data found, e.g. notes and receipts from general practitioners, as well as simply older data. While there have been considerable efforts been spent on the topic of structuring this semi- and unstructured data vaults during the last years (e.g. by Heurix in [39] and [40]), a comparison of this research with a subsequent identification of the major weaknesses and research gaps is needed. Following this basic analysis, research into constructing a working mechanism needs to be conducted.

Practical Considerations. In order to spread these anonymization techniques, efficient implementations are needed, preferably open source in order to enable the researchers to optimize the algorithms with respect to the data. An example for a framework can be found in [41], with an outline for an optimization for biomedical datasets in [42]. This also includes research on the optimization of the algorithms, which e.g. has been conducted by El Emam et. al in [43] for basic k-anonymity. Furthermore, for review processes, an interface allowing the anonymous exchange of data sets in the course of the peer-review process would be a valuable addition.

6.4 Influencing Other Research Areas

In other research areas, the problem of privacy in data-driven science is rather new which results in a striking absence of specific laws or regulations. In classical IT security we are under the impression that currently research data is usually held back instead of released in an anonymized state. In our opinion this is largely due to a lack of rules for anonymous data publishing which pushes responsibility for privacy protection onto individual researchers, thus resulting in uncertainty. This poses a major problem for the reproducibility of results, which is one of the main pillars of modern science and it's underlying review paradigm. Furthermore, todays anonymization concepts mostly come from medical research and are therefore designed for highly structured data, thus often cannot be used for other types of data. This again opens up a new area for research into new methods. In view of the rapidly growing trend towards more data-driven research, these new methods will be needed rather sooner than later. Specific policies and guidelines governing the public availability of data in (data-driven) science would also be helpful in order to guarantee the validity of published research.

7 Conclusion

In this chapter, we discussed several important anonymization operations and models for ensuring privacy, especially relating to the medical sector, focussing on data-driven medical research. Anonymization methods aim to obfuscate the identity of the record owners, taking into account not only direct identifiers but also quasi-identifiers. The eligible anonymization methods depend on the internal structure of data and its external representation. Once the appropriate method has been identified, the requirements of the chosen privacy models can be satisfied, but each time a data set is anonymized, information is lost or the data may be distorted. Furthermore, we outlined several interesting research questions that need to be tackled in order to heighten the security margin for protecting privacy, as well as produce more significant anonymized data sets for analysis.

On related terms, an even more general problem concerning the reproducibility of research, lies in the authenticity of the used data. While this does not directly relate to the topic of anonymization as discussed in this work, it is vital to take into account that the data used in data-driven medical science must be trustworthy, may it be anonymized or not. Still, anonymization can hinder the inspection and/or validation of data, thus we see additional research questions arising from this antagonism of protecting privacy on the one side and providing means of validating data on the other. Furthermore, researchers in data driven-science must always have in mind that the proper validation of their data with respect to originality and authenticity, as well as of the algorithms in use is of the utmost importance.

Acknowledgments. The research was funded by COMET K1 and grant 840824 (KIRAS) by the FFG - Austrian Research Promotion Agency.

References

1. Chawla, N.V., Davis, D.A.: Bringing big data to personalized healthcare: A patient-centered framework. Journal of General Internal Medicine 28, S660–S665
2. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
3. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinformatics 15(suppl. 6), I1 (2014)

4. Emmert-Streib, F., de Matos Simoes, R., Glazko, G., McDade, S., Haibe-Kains, B., Holzinger, A., Dehmer, M., Campbell, F.: Functional and genetic analysis of the colon cancer network. BMC Bioinformatics 15(suppl. 6), S6 (2014)
5. Jacobs, A.: The pathologies of big data. Communications of the ACM 52(8), 36–44 (2009)
6. Craig, T., Ludloff, M.E.: Privacy and Big Data: The Players, Regulators and Stakeholders. Reilly Media, Inc., Beijing (2011)
7. Weippl, E., Holzinger, A., Tjoa, A.M.: Security aspects of ubiquitous computing in health care. Springer Elektrotechnik & Informationstechnik, e&i 123(4), 156–162 (2006)
8. Breivik, M., Hovland, G., From, P.J.: Trends in research and publication: Science 2.0 and open access. Modeling Identification and Control 30(3), 181–190 (2009)
9. Thompson, M., Heneghan, C.: Bmj open data campaign: We need to move the debate on open clinical trial data forward. British Medical Journal 345 (2012)
10. Hobel, H., Schrittwieser, S., Kieseberg, P., Weippl, E.: Privacy, Anonymity, Pseudonymity and Data Disclosure in Data-Driven Science (2013)
11. Bonneau, J.: The science of guessing: analyzing an anonymized corpus of 70 million passwords. In: 2012 IEEE Symposium on Security and Privacy (SP), pp. 538–552. IEEE (2012)
12. Chia, P.H., Yamamoto, Y., Asokan, N.: Is this app safe?: a large scale study on application permissions and risk signals. In: Proceedings of the 21st International Conference on World Wide Web, pp. 311–320. ACM (2012)
13. Dey, R., Jelveh, Z., Ross, K.: Facebook users have become much more private: A large-scale study. In: 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 346–352. IEEE (2012)
14. Siersdorfer, S., Chelaru, S., Nejdl, W., San Pedro, J.: How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In: Proceedings of the 19th International Conference on World Wide Web, pp. 891–900. ACM (2010)
15. West, R., Leskovec, J.: Human wayfinding in information networks. In: Proceedings of the 21st International Conference on World Wide Web, pp. 619–628. ACM (2012)
16. Zang, H., Bolot, J.: Anonymization of location data does not work: A large-scale measurement study. In: Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, pp. 145–156. ACM (2011)
17. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 571–588 (2002)
18. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
19. Fung, B., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR) 42(4), 14 (2010)
20. Dalenius, T.: Finding a needle in a haystack-or identifying anonymous census record. Journal of Official Statistics 2(3), 329–336 (1986)
21. Pfitzmann, A., Köhntopp, M.: Anonymity, unobservability, and pseudonymity - A proposal for terminology. In: Federrath, H. (ed.) Anonymity 2000. LNCS, vol. 2009, pp. 1–9. Springer, Heidelberg (2001)
22. Hobel, H., Heurix, J., Anjomshoaa, A., Weippl, E.: Towards security-enhanced and privacy-preserving mashup compositions. In: Janczewski, L.J., Wolfe, H.B., Shenoi, S. (eds.) SEC 2013. IFIP AICT, vol. 405, pp. 286–299. Springer, Heidelberg (2013)

23. Wang, K., Fung, B.C., Philip, S.Y.: Handicapping attacker's confidence: an alternative to k-anonymization. Knowledge and Information Systems 11(3), 345–368 (2007)
24. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 223–228. ACM (2004)
25. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM (2005)
26. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, pp. 25–25. IEEE (2006)
27. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 277–286. ACM (2006)
28. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 3 (2007)
29. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: ICDE, vol. 7, pp. 106–115 (2007)
30. Li, J., Tao, Y., Xiao, X.: Preservation of proximity privacy in publishing numerical sensitive data. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 473–486. ACM (2008)
31. Heurix, J., Karlinger, M., Neubauer, T.: Pseudonymization with metadata encryption for privacy-preserving searchable documents. In: 2012 45th Hawaii International Conference on System Science (HICSS), pp. 3011–3020. IEEE (2012)
32. Neubauer, T., Heurix, J.: A methodology for the pseudonymization of medical data. International Journal of Medical Informatics 80(3), 190–204 (2011)
33. Heurix, J., Neubauer, T.: Privacy-preserving storage and access of medical data through pseudonymization and encryption. In: Furnell, S., Lambrinoudakis, C., Pernul, G. (eds.) TrustBus 2011. LNCS, vol. 6863, pp. 186–197. Springer, Heidelberg (2011)
34. Noumeir, R., Lemay, A., Lina, J.M.: Pseudonymization of radiology data for research purposes. Journal of Digital Imaging 20(3), 284–295 (2007)
35. Agrawal, R., Kiernan, J.: Watermarking relational databases. In: Proceedings of the 28th International Conference on Very Large Data Bases, pp. 155–166. VLDB Endowment (2002)
36. Deshpande, A., Gadge, J.: New watermarking technique for relational databases. In: 2009 2nd International Conference on Emerging Trends in Engineering and Technology (ICETET), pp. 664–669 (2009)
37. Kieseberg, P., Schrittwieser, S., Mulazzani, M., Echizen, I., Weippl, E.: An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata. Electronic Markets - The International Journal on Networked Business (2014)
38. Schrittwieser, S., Kieseberg, P., Echizen, I., Wohlgemuth, S., Sonehara, N., Weippl, E.: An algorithm for k-anonymity-based fingerprinting. In: Shi, Y.Q., Kim, H.-J., Perez-Gonzalez, F. (eds.) IWDW 2011. LNCS, vol. 7128, pp. 439–452. Springer, Heidelberg (2012)
39. Heurix, J., Rella, A., Fenz, S., Neubauer, T.: Automated transformation of semi-structured text elements. In: AMCIS 2012 Proceedings, pp. 1–11 (August 2012)

40. Heurix, J., Rella, A., Fenz, S., Neubauer, T.: A rule-based transformation system for converting semi-structured medical documents. Health and Technology, 1–13 (March 2013)
41. Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., Kuhn, K.A.: Flash: efficient, stable and optimal k-anonymity. In: 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), 2012 International Confernece on Social Computing (SocialCom), pp. 708–717. IEEE (2012)
42. Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., Kuhn, K.A.: Highly efficient optimal k-anonymity for biomedical datasets. In: 2012 25th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. IEEE (2012)
43. El Emam, K., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.P., Walker, M., Chowdhury, S., Vaillancourt, R., et al.: A globally optimal k-anonymity method for the de-identification of health data. Journal of the American Medical Informatics Association 16(5), 670–682 (2009)