ShinyAnonymizer: A Tool for Anonymizing Health Data

Marios Vardalachakis[1,2], Haridimos Kondylakis[1,2][a], Lefteris Koumakis[1,2][b],
Angelina Kouroubali[1][c] and Dimitrios G. Katehakis[1,3][d]
[1]*Computational Biomedicine Laboratory, FORTH-ICS, Heraklion, Crete, Greece*
[2]*Department of Informatics Engineering, Technological Educational Institute of Crete, Greece*
[3]*Center for eHealth Applications and Services, FORTH-ICS, Heraklion, Crete, Greece*
mariosggg@gmail.com, {kondylak, koumakis, kouroub, katehaki}@ics.forth.gr

Keywords: Personal Health Data, Security, Privacy, Data Anonymization.

Abstract: Processing and managing sensitive health data requires a high standard of security and privacy measures to ensure that all ethical and legal requirements are respected. Data anonymization is one of the key technologies to this purpose. However, the plethora and the complexity of the available anonymization techniques make it difficult for a non-expert to select and apply the appropriate technique. In this paper, we report on Shiny Database Anonymizer, a tool enabling the easy and flexible anonymization of available health data, providing access to state of the art anonymization techniques, incorporating also multiple data analysis visualization paradigms. In addition, a number of encryption and hashing techniques are presented.

1 INTRODUCTION

In recent years, various health and medical institutions have collected a large quantity of medical data through their Electronic Health Records (EHRs) (Katehakis et al, 2011). If used effectively, these data can help acquire new knowledge for disease prevention, enabling also optimal decision-making (Schaller et al., 2016). Distributed health care systems consist of patient data and knowledge from health care institutions. Meaningful knowledge can be extracted from medical data which can be shared with other institutions to achieve knowledge interoperability among different heterogeneous health care systems (Kouroubali et al, 1997; Katehakis et al, 2017).

To this direction, data availability and sharing have become core elements in biomedical and healthcare research (Kondylakis et al., 2015d; Maniadi et al., 2013). However, patient data usually include sensitive information about personal history, diagnoses and medications (Kondylakis et al., 2015a). Multiple regulations and laws have been established worldwide in order to protect the privacy of individuals. A recent example in the European Union is the General Data Protection Regulation (GDPR) that came into force in 25 May 2018. The GDPR regulates sensitive data access and establishes measures for ensuring the privacy of patients when sharing medical data. A key technology to ensure data privacy is data anonymization using available anonymization techniques. Anonymization of data is needed in order for third parties to use patient information without compromising confidentiality. The information is anonymized in such a way that the characteristics of the data remain while the anonymity of personal identify of the person is preserved. A typical use case is the knowledge extraction from medical data (Berman, 2002) and the corresponding secondary usage (Kondylakis et al, 2016). The first and most important step of clinical data mining is the seamless data integration (Potamias et al., 2005) in a way that privacy is preserved (Aggarwal and Philip, 2008).

While anonymization is an important process for protecting privacy, the available tools allowing non-expert users to anonymize available data are limited. These tools are often not user-friendly enough while

[a] https://orcid.org/0000-0002-9917-4486
[b] https://orcid.org/0000-0002-8442-4630
[c] https://orcid.org/0000-0002-3023-8242
[d] https://orcid.org/0000-0002-3763-191X

Vardalachakis, M., Kondylakis, H., Koumakis, L., Kouroubali, A. and Katehakis, D.
ShinyAnonymizer: A Tool for Anonymizing Health Data.
In *Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2019)*, pages 325-332
ISBN: 978-989-758-368-1

they require a deep understanding of the corresponding algorithms. These facts limit their practical application. This paper focuses on current research activities related to the implementation of a tool for Anonymizing Health Data called ShinyAnonymizer. ShinyAnonymizer is able to connect to various databases, enabling non-expert users to easily select data from remote databases and then by using a point and click graphical interface, to anonymize the data with a plethora of available methods.

ShinyAnonymizer offers a wide range of available algorithms, explained for non-expert users in a non-technical language using examples, ensuring the usage of state of the art confidentiality models that mitigate attacks and privacy breaches. More specifically, ShinyAnonymizer supports a combination of the following privacy models:

- Anonymization privacy models (removing information, suppression, generalization, bottom and top coding)
- Hashing privacy models (md5, crc32, sha512, xxhash64)
- Encryption privacy models (Des, X-Des, Aes-512, blowfish)

In addition to multiple methods for privacy, the system also provides multiple data analysis visualization paradigms and statistics such as pie charts, bar charts, area charts, histograms and scatter plots.

ShinyAnonymizer uses well known and high productive anonymization algorithms and implements a carefully selected set of techniques that can handle a large spectrum of data anonymization tasks while being productive and easy to understand.

Finally, ShinyAnonymizer provides a stand-alone software library that can be easilty integrated and used into other systems. In addition, it is extendable to many more algorithms, is well-tested and carefully documented. Therefore, ShinyAnonymizer provides a robust environment for developing novel privacy models.

The whole platform has been developed in the context of iManageCancer EU project (http://imanagecancer.eu/) for empowering cancer patients. The project aims to build an advanced, standards-based and scalable semantic integration environment (Kondylakis, et al., 2017), enabling seamless, secure and consistent bi-directional linking of clinical research and clinical care systems, which among others, will empower patients to extract the relevant data out of the overwhelmingly large amounts of heterogeneous data and treatment information (Kondylakis, et al, 2015b; Kondylakis et al, 2015c). For enabling research organization to analyze health data of the patients willing to share their anonymized data, by signing the necessary consent form (Kondylakis, et al, 2017) the ShinyAnonymizer is used to preprocess the available data before enabling data analysis on them (Koumakis, et al, 2018). Fields like name, surname and address and other fields that could lead to leakage of personal information by reasonable means are anonymized, whereas data transmission among the various components of the framework are encrypted before being transferred.

The rest of this paper is structured as follows: Section 2 presents the ShinyAnonymizer workflow and architecture, describing in detail also the available components, and section 3 provides a walkthrough of the tool. A comparison with related tools is presented in Section 4. Finally, Section 5 concludes this paper and presents directions for future work.

2 SHINYANONYMIZER WORKFLOW AND SYSTEM ARCHITECTURE

2.1 Workflow

ShinyAnonymizer is targeted for domain experts who will anonymize data with the help of an IT person. The workflow consists of four steps as shown in Figure 1. The figure shows the process of anonymization and its alternatives. Our work aims at making data anonymization attractive to a broad range of end-users that might not have expertise in anonymization techniques. We decided to implement different algorithms that are intuitive to non-expert users and that can be easily configured.

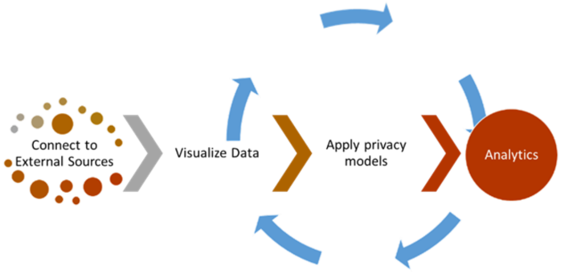


Figure 1: System Workflow.

The first step of the workflow involves connection, retrieval and visualization of already existing data stored in external databases, excel and

CSV files and others. ShinyAnonymizer supports the connection to most of the existing relational databases and file formats, accessible using various scripts of the R programming language. In each case, only the appropriate connection parameters should be defined and then the ShinyAnonymizer is able to connect to the appropriate data source in order to retrieve and process the available data.

The second step is to visualize the available data, enabling the active exploration of the available information. The users can select the fields to be further processed and can experiment with the available privacy models.

In the third step, the users can select among many different algorithms for encryption, hashing and anonymization.

Finally in the fourth step, the users are able to perform several analytic tasks in order to check the information available after the application of the various privacy models.

2.2 System Architecture

In order for the aforementioned workflow to be executed, a three-layered architecture has been designed and implemented, consisting of the Graphical User Interface (GUI), the Privacy models and the Data Source Application Programming Interfaces (APIs), as depicted in Figure 2. Detailed description of each one of those layers is provided below.

2.2.1 The GUI

The top layer of the ShinyAnonymizer system is the graphical user interface which is the front-end of the whole system. It is a web application using multiple graphs and visualization techniques for enabling data exploration and for visualizing the result of the various privacy models that are applied to the data. This layer has been implemented using HTML, JavaScript and CSS over the Shiny R package (Team, 2014). A screenshot of the GUI is shown in Figure 3. After configuring the connection to an external data source, the data are transformed and stored in an internal PostgreSQL database. Then the user is presented with a tabular view of the data which enable the active exploration of the various data fields and the identification of the data to be anonymized, encrypted or hashed.

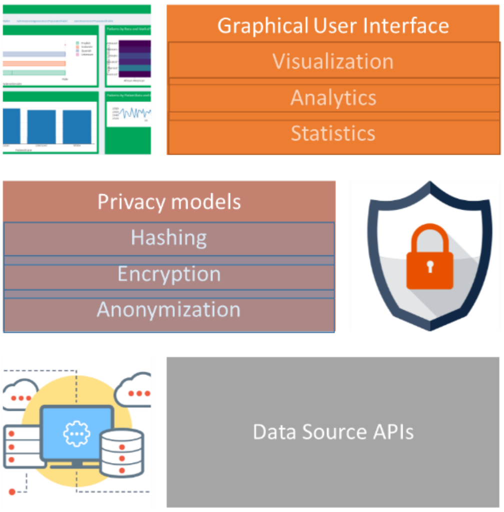
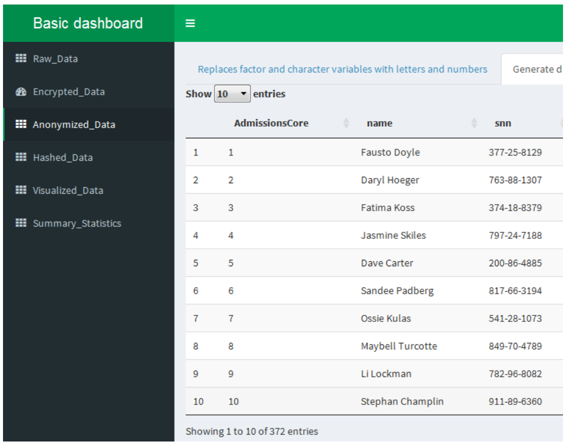


Figure 2: System Architecture.



Show 10 entries

	AdmissionsCore	name	snn
1	1	Fausto Doyle	377-25-8129
2	2	Daryl Hoeger	763-88-1307
3	3	Fatima Koss	374-18-8379
4	4	Jasmine Skiles	797-24-7188
5	5	Dave Carter	200-86-4885
6	6	Sandee Padberg	817-66-3194
7	7	Ossie Kulas	541-28-1073
8	8	Maybell Turcotte	849-70-4789
9	9	Li Lockman	782-96-8082
10	10	Stephan Champlin	911-89-6360

Showing 1 to 10 of 372 entries

Figure 3: The GUI of the system.

Before and after the application of the various privacy models, data exploration is available in tabular format and in the following visualization charts:

- **Pie Chart:** The pie chart shows the data as sectors from a circle and is accordingly useful for displaying data as total parts. Pie charts are separated into sections to present the values of different sizes.
- **Bar Chart:** The Bar chart displays the data differences in length according to their value. Usually it shows a comparison between several sets of data.
- **Histogram:** The histogram chart usually describes the data as continuous lines that pass through points defined by their items and values. As the histogram uses colors to desplay different areas, it is useful for emphasizing changes in values from several sets of similar data. A colored background helps visualize the differences clearly.
- **Area Chart:** Area charts are also called space graphs and are used to show graphical

representation of quantitative data. They display data in leap Areas.

- **Scatter Plot:** The scatter plot chart displays the data as points with coordinates and size determined by their item values. A bubble chart is useful for visualizing different scientific relationships.

Summary statistics are also presented to the user summarizing data ranges and diversity of the various columns, offering useful insights on the quality and the size of the available data.

2.2.2 The Privacy Models

In this layer, multiple algorithms are available for hashing encryption and anonymization. All algorithms have been implemented in R. Encryption is a technique of scrambling the data in order to remove sensitive information. The available encryption methods in our system are the following:

- **DES:** It is one of the most widely accepted publicly available cryptographic systems today. It uses a 56-bit key to encrypt the 64 bit block size data. It processes 64-bit inputs into 64-bit cipher-text and algorithm performs 16 iterations (Kaur et al, 2016).
- **X-DES:** is similar to DES. Its power can be increased using the Galois Fields. By increasing the size of the data structure and the length of the encryption key at least two times we can stop simple cryptographic attacks like brute force attacks etc. (Luminita et al, 2017).
- **AES-512:** has been one of the more powerful security protocols. This is because it uses higher length key sizes such as 128,192,256 and now 512 bits for encryption. This makes the algorithm more powerful to avoid cryptographic attacks and also it is faster than other state of the art algorithms available (Abidalrahman et al, 2011).
- **Blowfish:** As DES and Extended DES need significant processing power for execution, they have been replaced by the Blowfish algorithm. Blowfish algorithm is one of the fastest block ciphers in general use, except when changing keys. Each new key requires preprocessing equivalent to encrypting 4 kilobytes of text which is very slow compared to other block ciphers (Schneier, 1993).

Besides encrypting the available data, there are also methods for data anonymization. An anonymization algorithm use special operations to appropriately hide the personal identity of the available data. Multiple techniques have been implemented to this direction as well.

- **Fake Data:** This technique is used to derive massive amounts of fake data in the appropriate fields. Multiple functions are employed able to generate names, addresses, email, various dates, coordinates, telephone numbers etc. The end-user is able to select the specific field and the type of the random data to be generated.
- **Replace Characters by Random Letters and Numbers:** This technique is used with the purpose to protect selected fields by replacing factor and character fields by a combination of random sampled letters and numbers. Numeric columns can be transformed as well. The implemented algorithm generates 5-digits long character string by default for a selected column but it can be configured to generate larger codes as well.

Besides encrypting and anonymizing data, in many cases, it is important to guarantee that data reach their destination in their original form as sent by the sender or that they are coming from an authenticated/trusted source. Hashing functions are commonly used to this purpose. They can also be used as special mathematical functions that convert a numerical value into another compressed numerical value. The input to the hash function is an arbitrary length but the output is always of a fixed length. In addition an additional input is usually added to the plain message before hashing also known as salting to enhance the quality of the output. Values returned by a hash function are called message digest or simply hash values. Using a hash function, the input cannot be deciphered by knowing the hash function and the output, unless every possible input is tried. For this reason, it is considered a safe method for data protection. The available hashing methods in our system are the following:

- **MD5:** Generates a message digest of fixed 128 bits and it takes 64 rounds. It was created in 1992.The goal of MD5 is to produce digest that seems to be random. It is ordinarily used to audit the integrity of data and it is also introduced as a 32-digit hexadecimal number. Finally, it provides a fingerprint of a message of arbitrary length (RFC 1231, 1992).
- **SHA512:** MD5 is recent and common. That means that there exist a lot of methods for attacking MD5 (e.g. rainbow table attacks). SHA512 is similar to MD5 with one difference: SHA512 is better for security and has relative fewer attacks than MD5 (SHA-2, 2018).

- **CRC32:** CRC32 uses a fewer bits and can correct one bit errors. This is considered as an advantage against other algorithms for better protection of the data. Finally, the usage of systematical circular codes encodes the message by adding a fixed length check value for the scope of error detection (Peterson and Brown, 1961).
- **XXHASH64:** It is a fast hashing algorithm but writes only a few bytes (Templ et al, 2015).

It is important to note that the system is easily extensible to use more encryption, anonymization and hashing algorithms. It is only a matter of providing a new function call for enabling the usage of a new privacy model.

Besides hashing the available data, there are also methods for data anonymization. An anonymization algorithm uses special operations to appropriately hide the personal identity of the available data. Multiple techniques have been implemented to this direction as well.

- **Suppression:** It replaces a quality value with the special symbol "*".
- **Removing Information:** A quality value is replaced with a sequence of numbers that ranges for example from 1 to 469 (total number of observations).
- **Generalization:** This technique replaces quality values with semantically unvarying less special values hiding the details of attributes.
- **Bottom Coding:** It is an exposure constraint technique that includes limiting the minimum worth of a variable allowed on the file to prevent exposure of individuals or other units with extreme values in a delivery.

2.2.3 Data Source APIs

This layer includes multiple APIs foe enabling connection to external data sources such as relational databases and CSV and excels files. In addition, the appropriate REST function calls and the corresponding API are available for programmatically retrieving data from external data sources and applying multiple encryption, anonymization and hashing functions.

3 A WALKTHROUGH TO THE SHINYANONYMIZER

SA has been used for applying privacy models to iManageCancer patient data to properly anonymize them. However, for building the system and for demonstration purposes we are using a chimerical patient database found online (EmrBots, 2018) that consists of 100 patients, 372 admissions and 111,483 lab observations. These data are available as CSV file. By selecting the file location, automatically the data are imported into a PostgreSQL database and the different privacy models can now be applied and configured.

The GUI consists of six tabs that can be navigated using the top navigation bar. Table 1 gives an overview and description of the tabs in the GUI.

Table 1: Overview of screen tabs in the ShinyAnonymizer Application.

Tab	**Functionality**
Raw Data	Load and prepare the dataset
Encrypted Data	Encryption Methods
Anonymized Data	Anonymized Methods
Hashing Data	Hashing Methods
Visualized Data	Visualized techniques
Statistics Data	Statistics

The Raw Data tab provides basic information of the available data, showing the various tables and fields available and enabling a basic search over the data.

After loading the data, the user can navigate to the Encrypted Data tab to select the encryption methods to be applied. There are specific tabs for each one of the implemented algorithm (Blowfish, AES-512, DES and X-DES) and the user can select the column for the corresponding algorithm to be applied. The application for each algorithm has been preconfigured, however, the user can set further parameters (e.g. for the Blowfish Algorithm). For each case, short explanations are provided for the non-expert user, when clicking to the corresponding algorithm name. Through the Anonymized Data tab, the user is able to apply anonymization algorithms (fake data, random letters and numbers).

Hashing Algorithms can be used by selecting the Hashing Data tab. By selecting the Visualized Data tab, the available functions for data exploration, are available. As such pie charts, histograms, mosaic plots and some summary statistics can be used.

Finally, a summary of the available data is presented through various statistics on each field such as the number of patients, the min, median, mean and max for various fields, the types of the various fields etc.

4 COMPARISON WITH EXISTING WORKS

There are already many approaches providing data anonymization as open source tools. To our knowledge, the most well-known ones are the Sdcmicro, the Anonymizer and the ARX. These are described below.

Sdcmicro (Templ, et al., 2015) aims to protect the data from unauthorized users. It focuses on anonymizing data from statistical agencies and other institutions that are mostly confidential and can be used for the production of anonymized (micro) data, i.e. for the construction of public- and scientific-use files. In addition, various risk assessment methods are included. Sdcmicro borrows techniques from other fields. For example multivariate statistics are used to change or affect constant variables and to quantify information loss. In addition to the anonymization methods implemented in the Sdcmicro package, an inclusive set of risk and utility measures are also provided. Finally, it includes functions to measure, visualize and compare risk and utility throughout the anonymization process.

Another anonymization software, called Anonymizer (Hendricks, 2015) mostly permits users to rapidly and easily anonymize data containing Personally Identifiable Information (PII) through "convenience functions", i.e. using a combination of salting and hashing.

Table 2: The main characteristics of the open source anonymization tools.

	Anonymization Methods	**Hashing Methods**	**Encryption Methods**
SDCMicro	K-anonymity I-Diversity Randomization Adding noise Rank swapping recording	-	-
The Anonymizer	Fake data	SHA256 CRC32	-
ARX	K-anonymity I-diversity t-closeness differential privacy	-	-
Shiny Anonymizer	Fake Data Randomization	SHA512 MD5 CRC32 XXHASH64	DES X-DES AES-512 BLOWFISH

Yet another system for data anonymization is ARX (Abidalrahman, et al. 2011) aiming at providing scalability and usability. It supports also various anonymization techniques, methods for analyzing data quality and re-identification risks. In addition, it supports well-known anonymization methods such as k-anonymity, I-diversity, t-closeness and differential privacy. Table 2 elaborates on the key characteristics of those tools.

Similar to the aforementioned systems, we argue that multiple anonymization techniques should be combined and used as a single technique is not always optimal for a specific scenario. Anonymization, hashing and encryption models can be effectively and efficiently combined to ensure protection of the data from unauthorized usage. However when combining multiple state of the art algorithms, an optimal workflow should be provided with explanations and guidelines showing also examples of each transformation to the available data. Furthermore they should be complemented by at least a basic analytics framework enabling the active exploration of the available data. To the best of our knowledge ShinyAnonymizer (Shiny Anonymizer) is the only system enabling non-expert access to multiple privacy models, in a user-friendly manner, combining privacy with data exploration, offering multiple exploration options through a rich set of data analysis graphs.

5 DISCUSSION & CONCLUSION

This paper focused on the implementation of a tool called ShinyAnonymizer (Shiny Anonymizer) which offers multiple privacy models that can be applied on existing data, by non-expert users. The most well-known encryption, hashing and anonymization techniques have been implemented.

In addition, preliminary usage of the system within the iManageCancer project has shown that the system is reliable and can provide a vast amount of options to experts that are able to quickly select the appropriate fields to be anonymized. The tool is going to be released soon as open source in order to enable further real life usage.

However, having a tool like the one proposed might create a false sense of privacy, as eventually the domain experts should be trained and select the appropriate fields to be anonymized. To this task we envision that an IT expert will guide the domain expert at least the first times that the tool is used, explaining him the data view and the various options available.

As more technologies emerge like the block-chain, the internet of things and the artificial intelligent technologies, anonymization and encryption of data become more important than ever. Future work will aim on further development of the tool implementing new anonymization algorithms, testing different types of data, comparing the effectiveness of each implemented algorithm and linking the anonymization tool with the analytics framework (Koumakis, et al., 2018).

ACKNOWLEDGEMENTS

Work reported in this chapter was partially supported by the European Union's Horizon 2020 research and innovation programme under projects iManageCancer ("Empowering patients and strengthening self-management in cancer diseases", grant agreement No 643529, http://imanagecancer.eu), Bounce ("Predicting Effective Adaptation to Breast Cancer to Help Women to BOUNCE Back", grant agreement No 777167, https://www.bounce-project.eu), and MyPal ("Fostering Palliative Care of Adults and Children with Cancer through Advanced Patient Reported Outcome Systems", grant agreement No 825872, http://mypal-project.eu/).

REFERENCES

Abidalrahman, M., Jararweh, Y., Tawalbeh, L. (2011) AES-512: 512-bit Advanced Encryption Standard algorithm design and evaluation.*Information Assurance and Security* (IAS).

Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms. *In Privacy-preserving data mining* (pp. 11-52). Springer, Boston, MA.

Berman, J. J. (2002). Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine*, 26(1-2), 25-36.

EmrBots.Org (2018) Experiment with artificial large medical data-sets without worrying about privacy, http://www.emrbots.org/ [Online; accessed August 2018].

Hendricks, P. (2015) Anonymizer: Anonymize Data Containing Personally Identifiable Information. R package version 0.2.0. https://github.com/ paulhendricks/anonymizer [Online; accessed August 2018].

Katehakis, D. G., Halkiotis, S., & Kouroubali, A. (2011). Materialization of Regional Health Information Networks in Greece: Electronic Health Record Barriers & Enablers. Journal of Healthcare Engineering, 2(3), 389-403.

Katehakis, D. G., Kondylakis, H., Koumakis, L., Kouroubali, A., & Marias, K. (2017). Integrated Care Solutions for the Citizen: Personal Health Record Functional Models to Support Interoperability. EJBI, 13(1), 41-56.

Kaur, N., Sodhi, S. (2016) Data Encryption Standard Algorithm (DES) for Secure Data Transmission. *International Conference on Advances in Emerging Technology* (ICAET).

Kondylakis, H., Bucur, A., Dong, F., Renzi, C., Manfrinati, A., Graf, N., Hoffman, S., Koumakis, L., Pravettoni, G., Marias, K. and Tsiknakis, M., (2017, June). iManageCancer: Developing a platform for Empowering patients and strengthening self-management in cancer diseases. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems* (CBMS) (pp. 755-760). IEEE.

Kondylakis, H., Claerhout, B., Keyur, M., Koumakis, L., van Leeuwen, J., Marias, K., Perez-Rey, D., De Schepper, K., Tsiknakis, M., Bucur, A. (2016). The INTEGRATE project: Delivering solutions for efficient multi-centric clinical research and trials. *Journal of biomedical informatics*, 62, 32-47.

Kondylakis, H., Flouris, G., Fundulaki, I., Papakonstantinou, V., & Tsiknakis, M. (2015a). Flexible access to patient data through e-Consent. In *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, pp. 263-266.

Kondylakis, H., Koumakis, L., Hänold, S., Nwankwo, I., Forgó, N., Marias, K., Tsiknakis, M. and Graf, N., (2017). Donor's support tool: Enabling informed secondary use of patient's biomaterial and personal data. *International journal of medical informatics*, 97, pp.282-292.

Kondylakis, H., Koumakis, L., Kazantzaki, E., Chatzimina, M., Psaraki, M., Marias, K., Tsiknakis, M., (2015b). Patient Empowerment through Personal Medical Recommendations. *MedInfo*, 1117

Kondylakis, H., Koumakis, L., Psaraki, M., Troullinou, G., Chatzimina, M., Kazantzaki, E., ... & Tsiknakis, M. (2015c). Semantically-enabled Personal Medical Information Recommender. *In International Semantic Web Conference* (Posters & Demos).

Kondylakis, H., Spanakis, E. G., Sfakianakis, S., Sakkalis, V., Tsiknakis, M., Marias, K., ... & Dong, F. (2015d). Digital patient: personalized and translational data management through the MyHealthAvatar EU project. *IEEE EMBC*, pp. 1397-1400.

Koumakis, L., Kondylakis, H., Katehakis, D. G., Iatraki, G., Argyropaidas, P., Hatzimina, M., & Marias, K. (2018). A content-aware analytics framework for open health data. In *Precision Medicine Powered by pHealth and Connected Health* (pp. 59-64). Springer, Singapore.

Kouroubali, A., Starren, J., Barrows Jr, R. C., & Clayton, P. D. (1997). Practical lessons in remote connectivity. In Proceedings of the AMIA Annual Fall Symposium (p. 335). American Medical Informatics Association.

Luminița, S.,Mătăsaru, P.,Diaconu, F. (2017) Extended DES algorithm to Galois Fields. *IEEE Signals, Circuits and Systems* (ISSCS).

Maniadi, E., Kondylakis, H., Spanakis, E. G., Spanakis, M., Tsiknakis, M., Marias, K., & Dong, F. (2013, November). Designing a digital patient avatar in the context of the MyHealthAvatar project initiative. In *13th IEEE international conference on BioInformatics and BioEngineering*, pp. 1-4).

Peterson, W. W.; Brown, D. T. (1961) Cyclic Codes for Error Detection. *Proceedings of the IRE*. 49 (1): 228–235. doi:10.1109/JRPROC.1961.287814.

Potamias, G., Koumakis, L., &Moustakis, V. (2005). Mining XML Clinical Data: the HealthObs *System. Ingénierie des systèmesd'information*, 10(1), 59-79.

Prasser, F., et al. (2014) Arx-a comprehensive tool for anonymizing biomedical data. *AMIA Annual Symposium Proceedings*. Vol. 2014.pp. 984–993.

RFC 1321, (1992) The MD5 Message-Digest Algorithm.

Schaller, S., Marinova-Schmidt, V., Setzer, M., Kondylakis, H., Griebel, L., Sedlmayr, M., ... & Kolominsky-Rabas, P. L. (2016). Usefulness of a tailored ehealth service for informal caregivers and professionals in the dementia treatment and care setting: the eHealthMonitor Dementia Portal. JMIR research protocols, 5(2).

Schneier, B. (1993) Description of a new variable-length key, 64-bit block cipher (Blowfish*). International Workshop on Fast Software Encryption*. Springer, Berlin, Heidelberg.

SHA-2. https://en.wikipedia.org/wiki/SHA-2 [Online; accessed August 2018].

Team, R. C. (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. 2013.

Templ, M., Kowarik, A., Meindl, B. (2015) Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro,Journal of Statistical Software, Vol. 67 No.4, pp.1-36.

Collet, Y (2016) xxHash - Extremeley fast hash algorithm. https://github.com/Cyan4973/xxHash, [Online; accessed August 2018].