An Assessment of Data Guidelines in Cryopreservation Laboratories

Ana Fernandes
Universidade de Évora, Portugal
anavilafernandes@gmail.com

José Neves
Universidade do Minho, Portugal
jneves@di.uminho.pt

Margarida Figueiredo
Universidade de Évora, Portugal
mtf@uevora.pt

Henrique Vicente
Universidade de Évora, Portugal
hvicente@uevora.pt

ABSTRACT

On May 25, 2018, the General Data Protection Regulation (GDPR) entered into force in the European Union, which is of the utmost importance for monitoring its accomplishment by all organizations, especially those working in the health sector. However, it turns into a very difficult task, i.e., in order to meet this challenge, a practicable problem-solving methodology had to be developed and tested, that lead to a soft approach to computing using Artificial Neural Networks. On the other hand, the method chosen for data collection was the inquiry by questionnaire, in which 156 employees participated. The proposed system has an accuracy of about 90%, which can diagnose the fragility of the laboratory and encourage future improvements to ensure a high level of data protection.

CCS CONCEPTS

• **Security and privacy** → Human and societal aspects of security and privacy; Privacy protections; • **Computing methodologies** → Machine learning; Machine learning approaches; Neural networks.

KEYWORDS

Artificial neural networks, Cryopreservation laboratories, Data protection, General data protection regulation, Privacy and security

ACM Reference Format:

Ana Fernandes, Margarida Figueiredo, José Neves, and Henrique Vicente. 2021. An Assessment of Data Guidelines in Cryopreservation Laboratories. In *2021 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2021)*, February 19–21, 2021, Seoul, Republic of Korea. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3459104.3459200>

1 INTRODUCTION

New technologies have changed the economy and social life, and facilitated the free flow of personal data. The collection and dissemination of such data has increased significantly in all areas of society, not only because people are increasingly making their personal data available, but also because the new technologies enable the

use / dissemination of personal data by private / public institutions. Loss of control over personal data can cause physical, material or immaterial damage to citizens. Therefore, the issues of data protection have become increasingly important, and the General Data Protection Regulation (GDPR) in force in the European Union since May 25, 2018 was the last step in the direction of data protection [1]. In the health sector, given the extremely sensitive nature of the data, the application of this regulation is compulsory in order to ensure proper security and confidentiality [2]. In this sector, personal data can be associated with physical or mental health; past, present or future state of health; information from analyzes or examinations of a part of the body or samples; genetic data; Information about an illness, disability or risk of illness; anamnesis, treatments; or physiological/biomedical status. Indeed, the GDPR is still a controversial issue in the health sector and has been the subject of various studies and debates around the world [3, 4]. Demotes-Mainard et al. discuss the problem of transparency and data sharing (i.e. the foundations of open science), the potential of big data analytics, and the reuse of hospital data and health databases in the face of GDPR restrictions [5]. The authors stress that European countries should avoid poorly interoperable regulatory systems and make the necessary adjustments to protect users' data while streamlining clinical research. To this end, they emphasize the need to define in the context of health what is understood as anonymous data, pseudonym data as well as the public interest. In addition, the authors emphasize the importance of the patient's initial consent encompassing future research, and express how the data will be accessed, rather than restricting its use in future research. They also underscore the need for a technology infrastructure that enables researchers to access repositories and stores to reuse data [5]. Flaumenhaft and Ben-Assuli discuss the main barriers that have led to relatively low acceptance of personal health records and avoided their widespread use [6]. The authors state that concerns about the security and privacy of this sensitive data are one of the main obstacles. The study examines in detail the more relevant characteristics related to the security and privacy of health data in five jurisdictions. The authors point out that the revised legislation needs to highlight the GDPR as it is the most understanding and strictest data protection measures [6]. In some key areas, however, there are still different interpretations and show a certain ambiguity (e.g. the introduction of the term pseudonymization) [7]. Mense and Blobel compare the key features of the GDPR and the functional model of the HL7 Personal Health Record System (HL7 PHR-S FM). According to the authors, the security and data protection standards HL7 PHR-S FM

This work is licensed under a Creative Commons Attribution International 4.0 License.

ISEEIE 2021, February 19–21, 2021, Seoul, Republic of Korea
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8983-9/21/02.
<https://doi.org/10.1145/3459104.3459200>

enable the efficient implementation of the GDPR, i.e., meeting the goals of the legislature and clearly defining the steps organizations should follow to ensure adequate data protection [8]. This study was performed in cryopreservation laboratories, which deal with the storage of stem cells from both blood and umbilical cord tissue. In this type of activity, it is necessary to collect and process personal data from customers or potential customers. The aim of this study is therefore to check whether the laboratory can implement and comply with the GDPR. This paper consists of five sections. After a brief introduction to the problem, the basics used in this work are discussed, namely the Knowledge Discovery in Databases (KDD) process [9, 10] and a computer approach based on Artificial Neural Networks [11]. Section 3 presents the methodology, whereas in section 4 the results are presented and discussed. Finally, conclusions are drawn and future work is outlined.

2 FUNDAMENTALS

Technological means have provided an exponential growth, both in number of records and in complexity, regarding data storage. As a result of this effective increase in information, its processing through traditional methods has become increasingly difficult and complex. In this way, applications aimed to the task of Knowledge Discovery in Databases (KDD) have emerged, incorporating Data Mining (DM) tools [9, 10]. The KDD process involves several steps, namely selection, data pre-processing, data transformation, data mining and interpretation [9, 10]. The DM stage consists in choosing and using the methods and techniques that best fit the fulfilment of the established objectives [9].

2.1 Artificial Neural Networks

Data analysis is not a recent subject. For several years it has been carried out using mainly statistical methods. However, from an early stage, it became clear that the human brain analyses data and treats information differently, using learning processes [9]. ANNs were inspired by the nervous system of the human being and have been progressively applied in DM [11]. An ANN is a set of simple processing elements, called artificial neurons or nodes, organized in a highly interconnected parallel structure. They are similar to the behavior of the brain because, on one hand, knowledge is acquired from an environment through learning processes and, on the other hand, because knowledge is stored in the connections between the nodes [11].

From a historical point of view, ANNs had their origin in the 40s of the 20th century, with the work of Warren McCulloch and Walter Pitts. These authors presented a simplified model of the neuron (called an artificial neuron or node), based on the fact that the neuron is active or inactive, at a given moment, which corresponds to the true/false of the propositional logic or the one/zero of the Boolean algebra [12]. Other contributions followed, especially that of Rosenblatt, who introduced the perceptron model [13], which originated the most used network architectures, called the Multi-Layer Perceptron (MLP). These are organized by layers and the connections always propagate in one direction, with no cycles. During the last few years MLP were used to capture non-linear relationships among variables, in different areas, such as law

[14], water quality [15, 16], psychosocial risks management [17], or health [18, 19].

3 METHODS

This study was carried out in cryopreservation laboratories located in the north of Portugal. The age of the participants ranged from 22 to 69 years (average age 41 ± 19 years), with 57% women and 43% men. A questionnaire to assess the implementation of GDPR in cryopreservation laboratories was created and used for a cohort of 156 employees. To avoid the potential hidden errors associated to ostensibly random sampling methods that can lead to biased results [20], the questionnaires were applied to all employees of the laboratories. The questionnaire was divided into two sections, the first containing general questions (e.g., age, gender, academic qualifications and departmental areas of employees), while in the second the participants were asked to mark the option that best complete each statement according to their opinion. In the first section the answers are descriptive, whereas in the second one a Likert scale with four levels (i.e., *very reduced*, *reduced*, *medium* and *high*) was used.

The statements under consideration were organized into three groups, namely *Awareness Related Statements*, *Priority Related Statements*, and *Processes and Technologies Related Statements*. The former one comprises the statements, viz.

- The awareness of the laboratory on the guarantees of the rights of the holders of personal data is . . . ;
- The awareness of the laboratory on the obligations and principles of the GDPR is . . . ;
- The awareness of the laboratory on the general impact of the GDPR is . . . ; and
- The awareness of the laboratory regarding the implications of non-compliance with the GDPR is . . .

The second group encompasses the statements, viz.

- The management's priority in triggering the resources needed to implement the GDPR are . . . ;
- The priority of the information management department regarding the implementation of a data protection system is . . . ; and
- The priority of the group to which the laboratory belongs in unleashing the necessary resources for the implementation of the GDPR is . . .

Finally, the third one includes the statements, viz.

- The processes and technologies that guarantee the exercise of all the rights of data holders are . . . ;
- The information security management system that ensures a level of security appropriate to the data holders is . . . ; and
- The competence and training of the person who performs the functions of data protection officer are . . .

Pursuing the transformation of the qualitative data (collected using the questionnaire) into a quantitative one, the method suggested by Fernandes et al. was adopted [21]. Thus the set of n statements respecting to a particular theme is itemized into a unitary area circle split into n slices, where the marks in the axis correspond to each one of the possible options, corresponding the quantitative value to the total area, as specified below (section 4.3).

Figure 1: Frequency of answer to the statements included in the second section of the questionnaire.

The Waikato Environment for Knowledge Analysis (WEKA) was used to implement ANNs, maintaining the standard software parameters [22]. In each simulation, the database was randomly divided into two mutually exclusive partitions, i.e., the training and test sets.

4 RESULTS AND DISCUSSION

4.1 Sample Characterization

Applicants' age was categorized into age groups, i.e., less than 20 years of age, 20-30, 31-50, 51-65 and higher than 65 years of age. 64.8% of applicants are aged between 31 and 65 years old, 33.3% are lesser than 30 years of age, while 1.9% are higher than 65 years of age. Concerning academic qualification, 25.6% of the applicants expressed to have basic education, 23.7% declared that concluded secondary education, 40.4% stated to have finished a degree and 10.3% affirmed to have post graduate education.

4.2 Answer Frequency Analysis

Figure 1 displays the frequency of answering to the second part of the questionnaire, where applicants selected the option that best complete each statement according to their opinion. The statements S1 to S4 refer to the *Awareness*, S5 to S7 are related to *Priority*, and S8 to S10 are relative to *Processes and Technologies*. The analysis of results presented in figure 1 shows that for all the statements, regardless the dimension considered (i.e., *Awareness*, *Priority* and *Processes and Technologies*), most of the applicants (ranging between 89.1% and 94.2%) ticked the options *Medium* or *High* considering that the commitment of the laboratory regarding the implementation and fulfill of the GDPR is appropriate. In fact, only a small percentage of applicants, ranged between 5.8% and 10.9%, considers the efforts in the implementation of the mentioned guidelines unsatisfactory.

Bearing in mind the percentage of applicants which stated that the implementation of the GDPR is deficient, it would be interesting

to apply a new questionnaire in order to clarify which are the main barriers to the effective implementation of the guidelines. Such study would allow to compare the results obtained in the studied laboratories with the results published in the literature [6-8].

4.3 GDPR Implementation Assessment

To create a decision support system to assess the implementation of the GDPR, the results obtained in the second section of the questionnaire were used to train and test an ANN. Bearing in mind that the data collected has a qualitative nature, it was necessary to quantify them. For this purpose, the method suggested by Fernandes et al [21] was chosen. Aiming to exemplify the procedures, figure 2 presents the answers of applicant 1 to the mentioned section of the questionnaire.

For each statement group (i.e., *awareness*, *priority*, and *processes and technologies*) the answers were itemized into a unitary area circle. The marks in the axis correspond to each alternative, i.e., *very reduced*, *reduced*, *medium* and *high*. Exemplifying with the group of statements regarding *awareness*, the answer to S1 was *high* and the correspondent area is given by $\frac{1}{4} \times \pi \times (\frac{1}{\sqrt{\pi}})^2 = 0.25$, in S2 and S4 were marked the option *medium* and the areas are $\frac{1}{4} \times \pi \times (\frac{3}{4} \times \frac{1}{\sqrt{\pi}})^2 = 0.14$. Finally, for S3 the answer was *reduced* and the area is $\frac{1}{4} \times \pi \times (\frac{2}{4} \times \frac{1}{\sqrt{\pi}})^2 = 0.06$. The total area (i.e., 0.59) is the sum of the partial ones, being the quantitative value regarding the *awareness* group for applicant 1 (figure 3). Proceeding in a similar way for the remaining statements groups (i.e., *priority* and *processes and technologies*) and for each of the 156 applicants, the results displayed in table 1 were obtained.

To obtain the best ANN model to assess the GDPR implementation different network structures have been elaborated and evaluated. The performance of ANN models was compared using the confusion matrixes [23]. The 3-3-2-1 topology (figure 4) was the one that presented the best performance in terms of accuracy and was selected to evaluate the implementation of the GDPR. The

Table 1: A fragment of the knowledge base for the GDPR implementation assessment.

Applicant	Awareness	Priority	Processes and Technologies
1	0.59	0.71	0.29
2	0.78	0.56	0.50
...	...	...	...
156	0.81	1.0	0.42

IMPLEMENTATION OF THE GENERAL DATA PROTECTION REGULATION (GDPR) QUESTIONNAIRE

PART II

Tick the option that best complete each statement according to your opinion.

1. Awareness Related Statements

	Very Reduced	Reduced	Medium	High
S1. The awareness of the laboratory on the guarantees of the rights of the holders of personal data is ...	☐	☐	☐	☒
S2. The awareness of the laboratory on the obligations and principles of the GDPR is ...	☐	☐	☒	☐
S3. The awareness of the laboratory on the general impact of the GDPR is ...	☐	☒	☐	☐
S4. The awareness of the laboratory regarding the implications of non-compliance with the GDPR is ...	☐	☐	☒	☐

2. Priority Related Statements

	Very Reduced	Reduced	Medium	High
S5. The management's priority in triggering the resources needed to implement the GDPR are ...	☐	☐	☒	☐
S6. The priority of the information management department regarding the implementation of a data protection system is ...	☐	☐	☐	☒
S7. The priority of the group to which the laboratory belongs in unleashing the necessary resources for the implementation of the GDPR is ...	☐	☐	☒	☐

3. Processes and Technologies Related Statements

	Very Reduced	Reduced	Medium	High
S8. The processes and technologies that guarantee the exercise of all the rights of data holders are ...	☐	☒	☐	☐
S9. The information security management system that ensures a level of security appropriate to the data holders is ...	☒	☐	☐	☐
S10. The competence and training of the person who performs the functions of data protection officer are ...	☐	☐	☒	☐

Figure 2: The answers of applicant 1 to the second section of the GDPR questionnaire.

Figure 3: The quantification process of the qualitative information collected in the second section of the GDPR questionnaire for applicant 1.

correspondent confusion matrix is shown in table 2 (the values displayed are the average ones, based on 25 experiments). Thus, it is possible to calculate the model accuracy for training set (92.5%, i.e., 99 correctly labeled in 107) and for test set (89.8%, i.e., 44 correctly labeled in 49). In this way, the assessment of GDPR implementation

Figure 4: A schematic view of the ANN selected for GDPR implementation assessment.

with the proposed ANN model is satisfactory, reaching accuracies of around 90%.

The focus on data acquisition should be concentrated on the more important variables, taking into account the model accuracy, depreciating or setting aside the matter least ones. Sensitivity analysis is related with model output response to variations in its input variables. It is a basic procedure that may be carried out after the modelling phase, and examines the model responses when the inputs are modified. Sensitivity according to variance [24] was used to compute the relative importance of the input variables. The results are shown in figure 5 and seem to indicate that ANN inputs affect the outputs in an analogous way, although the fluctuations on *Processes and Technologies Related Statements* show a slightly more pronounced effect. Thus, the organization should pay attention to all these factors for an effective implementation of the GDPR. These outcomes are corroborated by the results shown in figure 3. In fact, the 3 groups of statements show a similar frequency of negative responses (i.e., *very reduced* and *reduced*) and, consequently, a small variation in this type of answers should have impact on the output.

5 CONCLUSIONS AND FUTURE WORK

Privacy and data protection are a very old topic, but at the same time very current, being the subject of recent developments that promise to have a significant impact on the provision of health services with the entry into force of the GDPR. Ensuring data protection should be the responsibility not only of citizens, but also of health professionals/technicians in the course of their activities. However, it is difficult to assess the level of GDPR implementation since it is a matter that deals with different variables with complex relationships among them. Therefore, and in order to assess

Table 2: Confusion matrix regarding ANN model for the GDPR implementation assessment.

Target	Predictive					
	Training Set			Test Set		
	Reduced	Medium	High	Reduced	Medium	High
Reduced	11	0	0	5	0	0
Medium	5	67	0	3	31	0
High	0	3	21	0	2	8

Figure 5: Relative importance of the input variables of the ANN selected for GDPR implementation assessment.

the level of implementation of the GDPR in a cryopreservation laboratory, a data acquisition and evaluation framework was developed and experienced in practice. The emphasis was putted on processing of information, being the data gathered using inquiry by questionnaire. Complementary, this paper also propose an intelligent decision support system to appraise the level of GDPR implementation based on the ANN paradigm. This approach exhibits a satisfactory effectiveness, showing an accuracy of about 90%. Furthermore, it allows to recognize the fragilities of the laboratory, and help the decision makers to promote future improvements to ensure high levels of data protection. Future work will consider new factors, namely the training given to the employees on this matter or the volume of information processed. In addition, it is intended to extend the study to a larger sample in order to study the influence of other variables such as age, gender or academic qualifications of employees on their perception about the *GDPR* implementation.

ACKNOWLEDGMENTS

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

REFERENCES

- [1] European Parliament, *EU General Data Protection Regulation – Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* (European Parliament, Brussels, 2016)
- [2] K. Sørensen, S. Van den Broucke, J. Fullam, G. Doyle, J. Pelikan, Z. Slonska, H. Brand, *BMC Public Health* **12**, 13 pp. (2012)
- [3] C. Morriss-Roberts, K. Oulton, D. Sell, J. Wray, F. Gibson, *Lancet Child Adolesc. Health* **2**, 696 (2018)
- [4] B. McCall, *Lancet* **391**, 1249 (2018)
- [5] J. Demotes-Mainard, C. Cornu, A. Guérin, P.-H. Bertoye, R. Boidin, S. Bureau, J.-M. Chrétien, C. Delval, D. Deplanque, C. Dubray, L. Duchossoy, V. Edel, C. Fouret, A. Galaup, F. Lesaulnier, M. Matei, F. Naudet, V. Plattner, M. Rubio, F. Ruiz, S. Sénéchal-Cohen, T. Simon, A. Vidal, A. Viola, M. Violleau, *Thérapies* **74**, 31 (2019)
- [6] Y. Flaumenhaft, O. Ben-Assuli, *Health Policy* **122**, 815 (2018)
- [7] M. Mourby, E. Mackey, M. Elliot, H. Gowans, S. E. Wallace, J. Bell, H. Smith, S. Aidinlis, J. Kaye, *Comput. Law Secur. Rev.* **34**, 222 (2018)
- [8] A. Mense, B. Blobel, *Eur. J. Biomed. Inform.* **13**, 27 (2017)
- [9] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques* (Morgan Kaufmann, San Francisco, 2012)
- [10] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, *Data Mining – Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, Cambridge, 2017)
- [11] S. Haykin, *Neural Networks and Learning Machines* (Prentice Hall, New York, 2009)
- [12] W. S. McCulloch, W. Pitts, *Bull. Math. Biophys.* **5**, 115 (1943)
- [13] F. Rosenblatt, *Psychol. Rev.* **65**, 386 (1958)
- [14] T. S. Nguyen, L. M. Nguyen, S. Tojo, K. Satoh, A. Shimazu, *Artif. Intell. Law* **26**, 169 (2018)
- [15] G. B. Ruben, K. Zhang, H. Bao, X. Mao, *Water Resour. Manage.* **32**, 273 (2018)
- [16] A. Fernandes, H. Chaves, R. Lima, J. Neves, H. Vicente, *IOP Conf. Ser.: Earth Environ. Sci.* (to be published)
- [17] A. Fernandes, M. Figueiredo, J. Ribeiro, J. Neves, H. Vicente, *Saf. Health Work-KR* doi: 10.1016/j.shaw.2020.07.003
- [18] N. Shahid, T. Rappon, W. Berta, *PLoS ONE* **14** (2), e0212356 (2019)
- [19] S. Ghosh, B. P. Chattopadhyay, R. M. Roy, J. Mukherjee, M. Mahadevappa, *Artif. Intell. Med.* **96**, 45 (2019)
- [20] Y. Shang, *IEEE Trans. Syst. Man Cybern. Syst.* **49**, 821 (2017)
- [21] A. Fernandes, H. Vicente, M. Figueiredo, M. Neves, J. Neves, *Lect. Notes Comput. Sci.* **10018**, 415 (2016)
- [22] E. Frank, M. Hall, I. H. Witten, *The WEKA Workbench – Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques” Morgan Kaufmann, Fourth Edition, 2016*. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf
- [23] J. Vilhena, H. Vicente, M. R. Martins, J. Grañeda, F. Caldeira, R. Gusmão, J. Neves, J. Neves, *International Journal of Reliable and Quality E-Healthcare* **6**, 48 (2017)
- [24] R. Kewley, M. Embrechts, C. Breneman, *IEEE Trans. Neural Netw.* **11**, 668 (2000)