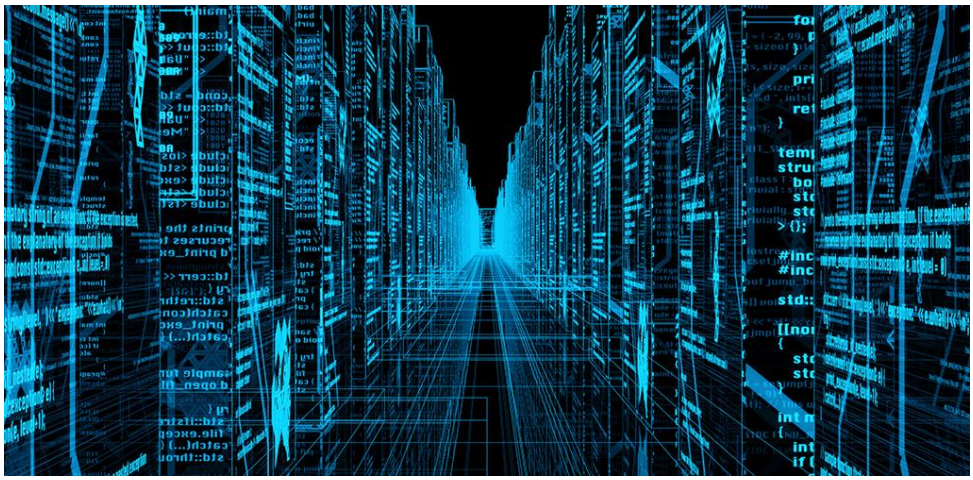
H E L L E N I C
N A T I O N A L B I O E T H I C S C O M M I S S I O N

R E P O R T

Big Data in Health

Rapporteurs: Takis Vidalis, Vasiliki Mollaki

Translation: Lydia Barba



HELLENIC NATIONAL BIOETHICS COMMISSION

Neofytou Vamva 6, P.C. 10674 Athens

Tel.: 0030 210- 88.47.700, Fax: 0030 210- 88.47.701

E-mail: secretariat@bioethics.gr, url: www.bioethics.gr

Contents

1. Introduction

"*Big data*" or else "*massive data*" constitute a collection of large and complicated sets of data which can lead to correlations and conclusions impossible to exist when analysis is restricted to individual data. Therefore, big data offer the possibility of finding solutions to difficult problems.

In the health sector, the use of big data is constantly expanding, so that epidemiological data can be collected, or complex diseases can be studied. The most usual sources of health data are electronic health record systems electronic systems for drug prescription, but also big databases used in research and clinical studies. However, information for everyday activities regarding the health of individuals can also be obtained by their posts on social media and the electronic files of their online transactions.

Although big data are important for several sectors, their dynamic in biomedicine, in both clinical practice and research, is outstanding. However, their use raises ethics issues, mostly because of the absence of consent from the data subjects to use their sensitive health information, because of the risk of information breach while storing and transferring data, and also because of the ambiguity regarding the ownership status of the data.

2. The actual facts

2.1. Features of big data

Big data are characterized by the so-called "4Vs+1": **Volume, Velocity, Variety, Veracity** and **Value.** That is, they are characterized by the large volume of information (which can be structured or not), the high velocity at which the data are collected and analyzed, the increased complexity and variety of data arriving at different formats, their reliability and the worth produced by their analysis.

The analysis of big data requires special statistical techniques and specialized information technology scientists, whereas, specifically in the process of big health data, the contribution of bioinformaticians is necessary.

The use and process of big data in health is, in most cases, carried out in anonymized or pseudoanonymized (or pseudonymized) data, so the privacy of the data subject can be protected and the processing purposes can be served. *Anonymized* are data that do not allow the determination or the identification of the data subject identity, even if they are combined with other data from any other source. Pseudonymization of data includes the replacement of any identification characteristics with a pseudoname (a code) or, in other words, a value that does not allow the direct definition of the data subject identity.

2.2. Possible data sources concerning health

During the last decades, developments are rapid regarding the available techniques of collection, sharing and analysis of large collections of information. The wide internet use, the creation of big databases and the development of portable devices has favored the collection of data. Indicative sources from where information can be obtained and analyzed are listed below:

a) **Scientific databases and study registries**, including clinical and laboratory datasets, that are either open and freely accessible online or they are closed and locally stored in labs. Particularly the volume of genetic data produced is expected to dramatically increase in databases, due to the development of new sequencing technologies and their reduced cost.[1] These databases include registries of clinical studies or trials in humans, which have been carried out in the past or are currently being carried out. Finally, these databases also include the international scientific literature.

b) **Clinical and administrative files**, and mainly electronic health records, which include epidemiological data, patients medical and family history but also possible therapeutic details and response to certain treatments.[2] The electronic drug-prescription system and insurance records Institutions are also significant health data sources.

c) **Portable devices and health applications**, including smart phones, smart watches and smart health bands, which are emerging monitoring technologies of personal health. [3,4]

d) **Non-scientific sources**, such as everyday online shopping, searches and posts of an individual on social media. A characteristic example of non-scientific sources is the case of Google Flu Trends. In 2008, Google researchers realized that there was an increasing number of searches for the flu, and by using the IP (Internet Protocol) address of each user, they could predict potential flu epidemics in several areas in the USA.[5] At certain time point, Google Flu Trends predictions coincided with those of the American Center for Disease Control and Prevention (CDC).[6] Other known examples of large-scale data use for predicting trends in public health are: Flu Trends,[5] the Twitter Influenza Surveillance[7] and the Facebook Emotional Contagion study.[8]

The above-mentioned programs/sites received a lot of criticism mostly because of the absence of consent of the data subjects, whose data were used for research. On the one hand, these examples highlight the opportunities provided by big data in health; on the other hand, they raise issues about the ethics of their use, the protection of private life, the privacy of health data and the autonomy of research participants. Even though these data are posted online on social media by the users themselves, they are collected and analyzed without their consent. The public is getting increasingly informed about the possible ways in which their data can be used, but often ignore the potential of these information.

2.3. Possible benefits of big data in health

In the health sector, big data could possibly improve decision-making, disease prevention and spread of diseases, but also reduce the cost of health care, because of their features (4Vs+1).

Big data provide the ability to doctors and researchers to collect a large volume of information and share it widely with other members of the scientific

community. Big data analysis can possibly facilitate the correlation of several parameters with diseases, such as the definition of new biomarkers which would predict -to a certain extend- whether a person will develop a diseased and when. This enables the investigation of complex and rare diseases in-depth and in a short period of time, which so far, were difficult to be studied with the analysis of small datasets.

Particularly, electronic health record datasets and their analysis provide the possibility to identify new risk factors, to define epidemiological facts, to estimate the life expectancy for various diseases, to record and prevent side effects, and even acknowledge and extenuate medical errors. In general, electronic health record and electronic drug-prescription systems, favor recording, sending and exchange of data among health professionals for clinical, epidemiological, administrative and research purposes. Consequently, the analysis of big data arising from electronic health record and electronic drug-prescription systems can help develop or reform healthcare policies which will need to be applied.[2]

In the literature, several examples of analysis of big health data have been described, with promising results about the benefits of big data. These studies included mostly the collection and analysis of data from electronic health records.[9,10,11,12,13]

In addition, big data analysis provides the ability of better monitoring of public health, combining geographical and social variables with health data.[14]

2.4. Possible risks of big data technology

Big data involve risks related to ethics issues, as discussed below (Section 3), several organizational issues but also technical or methodological difficulties in their processing.[15] Organizational and technical issues exceed the purpose of the present Report. However, it is critical to note the risk of de-identification of a person, i.e. the probability of data be combined with other datasets that may exist freely online or in other kinds of records, resulting in the identification of the data subject. The de-

identification risk is known to all scientists dealing especially with big data, and this is exactly why specific measures have been taken to mitigate it. The literature includes studies that show that the anonymity or the pseudo-anonymity of data is not sufficient to prevent the de-identification.[16,17,18,19,20]

3. The ethics issues

There is no doubt that large databases in the health sector can play a significant role in the development of new drugs and medicotechnological products or methods for the treatment and prevention of disease, as well as the economic rationalization of health services and generally, the public health policy. However, these benefits seem to be mitigated by the risks that arise in terms of autonomy and privacy guarantees of the data subjects whose sensitive information are used. The harmonization of the value of health protection with the latter two values is, therefore, necessary.

The possibility of collecting a large volume of personal health data either to serve clinical applications or public health policies, or for research purposes, gives rise to a dual concern in terms of bioethics: a) in ensuring consent of data subjects, and b) in maintaining confidentiality.

3.1. The consent issue

Big data collections are created mostly by interconnection methods with existing files in units that maintain primary health data collections (hospitals, clinics, health centers, diagnostic centers, laboratories, insurance organizations, etc.). These primary collections are based on the explicit and specific consent of the data subjects. These individuals have been informed in advance for the purpose of collection and storage of data: therefore, their consent covers this particular purpose and no other.

So, if there is going to be an interconnection of primary collection files, in order to serve wider aims (beyond the aim to which each specific consent relates to),

the assurance of the person's autonomy presupposes, in principle, a new consensus adapted to the purposes of this interconnection. The practical problem here is that, as a rule, the "return" process to each subject of primary data is difficult, either because of the large volume that is intended to be collected, or because of the failure to locate people who had provided data in the past. If the problem is not detected, the creation of big databases is unfeasible.

A specific question is whether it would be possible to have an effective assurance of the person's autonomy by alternative ways beyond the standard, which requires specific and prior consent, to facilitate the secure interconnection of files for the benefit of serving wider purposes in Health and Research.

An alternative option is the creation of large data collections, based on the so-called "presumed consent" of the donors. This choice was adopted in the well-known case of establishing the Icelandic bio-bank, with the involvement of the state and a private company, causing reactions.[21] In the "presumed consent" system, health data subjects are presumed to have tacitly consented to any future use, retaining the right to withdraw their data from the collection whenever they wish (opt out). This way, it is considered that the interconnection of files and the processing of big data are facilitated, without prejudice to the data subject's autonomy and the individual control over sensitive information. The objection to presumed consent stresses the importance of prior expression of the individual's will (on the condition of appropriate information for specific use of personal data), as that is the only way that guarantees free choice and effective personal control. The reason is that, any "withdrawal" made later does not guarantee the control of use in the interim period, and it is up to the data subject's initiative to learn about possible uses which have already been launched.

The second alternative option is to abandon the requirement for a "specific" prior consent of the data subject, and the acceptance of a broader consent (covering for example more possible future uses). In this case, the advantage is the preservation of the condition of previous consent after being informed –hence of the actual control of the data subject. On the other hand, a disadvantage is both the

vagueness of the consent and the possible exclusion of future uses that cannot be predicted at the moment of consent.

An issue concerning consent also rises by the possibility of "incidental findings" during secondary use of data for research purposes. In particular, the question is whether researchers must inform the data subjects for research findings that come up during the course of research and may be important for the protection of the health of the latter, or if they need to seek special consent on this, taking into account the "right not-to-know".

The problem is known (and has been specifically addressed in the relevant Opinion issued by the HNBC), however, in the case of big data collections, the probability of discovering incidental findings is greatly increased and may even extend to anonymous initial data, which are not unlikely to be identified (see Section 3.2).

3.2. Assuring data confidentiality

The large volume of data resulting from the interconnection of more files poses a problem as to the maintenance of their confidential use. There are two specific problems here.

The first problem is related to the proportional increase in the possibility of breach as the size of the collection increases. The management of the latter often requires the involvement of more managers (technical and administrative) and involves a great deal of complexity in the confidentiality safeguards. This element, alone, increases the risks of violating the latter.

The second problem is related to relativizing the free use of anonymous data in case we have large collections. Anonymous data -i.e. data for which the identity of the data subject or the code to their anonymity is not known to the processor- are currently considered completely safe in terms of confidentiality, as is not technically possible to correlate them with other personal data. This element seems to be overturned in big data collections, as modern "deep mining" technologies lead,

under conditions, to the identification of anonymous data, as the volume of the collection provides many parameters that facilitate correlations for that purpose. Ultimately, big data appear to cancel the category of anonymous data, putting the issue of expanding the systems that assure confidentiality to the anonymous data.

4. The relevant legislation

Specific international, European or national legislation to regulate the establishment and operation of big databases do not exist up to this point. Neither the current legislation on the protection of personal data (Law 2472/1997), nor the new EU regulation 2016/679 (soon to be implemented)[22] include specific provisions.

Therefore, in principle, with regard to the establishment of such databases containing health data, the provisions on the protection of sensitive personal data, apply. According to it (art. 7 of Law 2472/1997), the interconnection of health data files:

- Requires authorization of the Data Protection Authority.
- Does not necessarily require consent of the data subjects.
- A necessary condition, if the purpose of interconnection concerns health issues, is that the processing is carried out by a person subjected to a duty of confidentiality or a relevant code of conduct, and
- a necessary condition, if the purpose of the interconnection concerns research, is the anonymization of the data.

It must be noted that the extension of medical confidentiality to third persons supporting the physician and are under his/her supervision (according to art. 13 par. 2a of Law 3418/2015) cannot be applied in the case of big databases. The reason is that the collection and processing of data is, in principle, carried out by scientific, technical and administrative personnel who do not have the status of a treating physician for the particular data subject. This means that, the data subjects' consent to the interconnection of the data files, remains as a necessary condition

that accompanies the condition of prior authorization by the Data Protection Authority.

The new EU Regulation (GDPR), nevertheless, creates a new picture for the use of data in research, which has a major impact on the establishment of big databases for research purposes in the field of bio-medicine. More specifically, the Regulation does not consider the specific prior consent of the data subject to be necessary, nor their anonymized processing, in contrast to the current Directive 95/44. The relevant provision of art. 9 of the GDPR leaves no doubt about this change, although in the preamble there is concern about the issue of consent and it appears that a broad (prior) consent is proposed rather than a specific consent which was known so far, so that, research is not hampered in exceptional cases .

This new provision is counterbalanced by the Regulation, on the one hand, by the right of the individual to withdraw his/her data (opt-out), on the basis of the general "right to be forgotten" and, on the other hand, by transferring the responsibility for the safe processing of data under conditions of confidentiality to the data controller and the new supervisor of data protection (Data Protection Officer). In addition, the Regulation explicitly mentions the Member State's power to provide stricter data protection rules.

Practically, this provision will certainly facilitate the use of data in research -especially their "secondary" use in projects other than these initially collected for, and the management of incidental findings, i.e. the ability to inform the data subject at any time, as their anonymity is no longer a necessary condition. On the other hand, the institutional control of data is practically limiting the possibility of withdrawing them, because it is indeed difficult for a data subject to be aware of any future use of data in a variety of research projects, unless he/she has been previously asked to provide specific consent.

To deprive completely a person from self-control over the use of sensitive health data creates a bioethical problem insofar as it affects the core of his/her personality; indeed a data subject does not actually have the opportunity to oppose the use of his/her data for the research purpose of which he/she may disagree (e.g.

for military purposes etc.). A single solution to this is that the individual national legislators should insist on additional safeguards, such as prior consent or anonymization of sensitive data.

The problem is getting more complicated if, in that case, national laws follow divergent policies, as stricter national regulations may discourage participation in international research projects, since the exchange of sensitive datasets will be objectively more difficult. So the national legislator must move here, seeking a balanced protection of personality and freedom of research.

5. Recommendations

Big data, whether they are health data or social or financial data, concern the core of the personality. For this reason Institutions or individuals (e.g. researchers) involved in the process, have a moral obligation to recognize the potential risks and to take appropriate measures to minimize potential damage to the subjects whose data are analyzed. For the responsible use of big data technology, all these involved must initially be aware that the data they are handling may have an impact on the data subjects. Then, specific safeguards should be taken to protect the confidentiality of information.

5.1. Policy of Institutions using big data technology

A necessary step before collecting and processing big health data from organizations, and education or research Institutes is the creation of a Research Ethics Committee with the responsibility to monitor ethics and examine project applications by approving or not access to large collections and their analysis.[23] It is also necessary to create an Ethics Code of conduct for the collection, access, management and processing of big data. Inter alia, the Ethics Code of conduct must provide the designation of a Data Protection Officer, to ensure that the Institution complies with data protection legislation.

5.2. Data protection measures

The use of technical methods for data protection is absolutely necessary. The minimum data protection measures prior to their analysis is their anonymization by removing identifiers from records such as name, home address, date of birth, telephone number, email, social security number, etc. Frequently, in order to avoid the identification of persons, the so called "k- anonymity" is used,[24] which imparts the property to every search of data so that the answers are limited to k-persons, with the aid of algorithms. In addition, noise is introduced into the required and encrypted algorithms, techniques which, in combination, ensure a relatively good level of security.

5.3. Informed consent

1. Non-anonymized data should be analyzed with the consent of the data subjects and the consent forms should include information on:
2. Data storage, who will have access to them and what measures will be taken to prevent access from third parties.
3. Possible data transfers between Institutions or researchers in the same or other countries, as well as whether data will be exchanged in their original form before processing occurs or just the results of their analysis.
4. The likelihood of producing incidental findings concerning the health of the data subjects and the policy to be followed in such cases. This is whether incidental findings will be communicated to the subject and which ones (e.g. only findings of clinical significance).[25]
5. The likelihood of commercial use of data.

BIBLIOGRAPHY

[1] Berkman BE, Shapiro ZE, Eckstein L, Pike ER. The Ethics of Large-Scale Genomic Research. In J. Collmann and S.A. Matei (eds.), Ethical Reasoning in Big Data, Computational Social Sciences, Springer International Publishing Switzerland 2016, pp 53-54.

[2] Vidalis T, Mollaki V. Electronic Health Record: Ethical Issues (2015). Report for the Hellenic National Bioethics Commission. Available at http://www.bioethics.gr/images/pdf/GNOMES/REPORT_ELECTRONIC_HEALTH_RECORD_EN_FINAL.pdf

[3] Mittelstadt BD, Fairweather NB, Shaw M, McBride N. The ethical implications of personal health monitoring. International Journal of Technoethics 2014, 5(2), 37–60.

[4] Boye N. Co-production of Health enabled by next generation personal health systems. Studies in health technology and informatics 2012, 177, 52–58.

[5] Ginsberg J, Mohebbi M, Patel R. Brammer L. Detecting influenza epidemics using search engine query data. Nature 2009, 457, 1012–1014.

[6] Ortiz J, Zhou H, Shay D, Neuzil K, Fowlkes A. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. PLoS One 2011, 6, e18687.

[7] Broniatowski DA, Paul MJ, Dredze M. Letter "Twitter: Big Data Opportunities," Science 2014, 345, 6193: 148.

[8] Kramer ADI., Guillory JE, Hancock JT. Experimental Evidence of Massive Scale Emotional Contagion through Social Networks. Proceedings of the National Academy of Sciences U.S.A. 2014, 111, 24 (2014): 8788-8790.

[9] Halamka J. Early experiences with big data at an academic medical center. Health Affairs 2014, 33(7): 1132–1138.

[10] Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet 2011 Jun;12(6):417-28.

[11] Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, Mani S, Levy MA, Dai Q, Denny JC. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc 2011;2011:1564-72.

[12] Weiss J, Natarajan S, Paissig P, McCarthy C, Page D. Machine learning for personalized medicine: Predicting primary MI from electronic medical records. AI Magazine 2012, 33(4): 33-45.

[13] Li L, Lu J, Xue W, Wang L, Zhai Y, Fan Z, Wu G, Fan F, Li J, Zhang C, Zhang Y, Zhao J. Target of obstructive sleep apnea syndrome merge lung cancer: based on big data platform. Oncotarget. 2017 Mar 28;8(13):21567-21578.

[14] Zhang X, Pérez-Stable EJ, Bourne PE, Peprah E, Duru OK, Breen N, Berrigan D, Wood F, Jackson JS, Wong DWS, Denny J. Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century. Ethn Dis 2017 Apr 20;27(2):95-106.

[15] Lipworth W, Mason PH, Kerridge I, Ioannidis JP. Ethics and Epistemology in Big Data Research. J Bioeth Inq. 2017 Mar 20.

[16] El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011, 6(12):e28071.

[17] Barbaro M, Zeller T, Hansell S. A face is exposed for AOL searcher no. 4417749. New York Times 2006 Aug 9;9.

[18] Cox J. 70,000 OkCupid Users Just Had Their Data Published. Motherboard. 12 May 2016. Available at http://motherboard.vice.com/read/70000-okcupid-users-just-had-their-data-published

[19] Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002, 10(05):557–70.

[20] Kloumann IM, Kleinberg JM. Community membership identification from small seed sets. InProceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014 Aug 24 (pp. 1366–1375). ACM.

[21] Gottweis H, Zatloukal K. Biobank Governance: Trends and Perspectives, Pathobiology 2007,74:206–211. Available at https://www.karger.com/Article/Pdf/104446

[22] Mitrou L. The General Data Protection Regulation, Sakkoulas 2017, Athens - Thessaloniki.

[23] Mollaki V, Vidalis T. Record of the Existing Regime Regarding the Establishment and Operation of Research Ethics Committees: Meeting with Representatives of Higher Education Institutions and Research Institutes (2015). Report for the Hellenic National Bioethics Commission (in Greek). Available at http://www.bioethics.gr/images/pdf/GNOMES/REPORT_ETHICS_COMMITTEES_FINAL_GR.pdf

[24] Latanya S, Pierangela S. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". Harvard Data Privacy Lab 1998.

[25] Mollaki V, Vidalis T. Incidental Findings in Research and Clinical Practice (2015). Report for the Hellenic National Bioethics Commission (in Greek). Available at http://www.bioethics.gr/images/pdf/GNOMES/REPORT_INCIDENTAL_FINDINGS_FINAL%20EN.pdf