

## Big Health Data: A systematic mapping study

S. LBRINI<sup>1,\*</sup>, A. FADIL<sup>2</sup>, H. RHINANE<sup>1</sup>, H. J. OULIDI<sup>2</sup>

<sup>1</sup>Geoscience Laboratory, Ain Chock Faculty of Science, Km 8 El Jadida Road, BP:5366, Casablanca, Morocco - (salmalbrini, h.rhinane@gmail.com)

<sup>2</sup>Hassania School of Public Works, km 7, El Jadida Road, BP: 8108, Casablanca, Morocco - (fadil.abdelhamid, hassane.jarar@gmail.com)

**KEY WORDS:** Big data, Health, real time, systematic mapping study, gis

### ABSTRACT:

The Big Data, a result of the digital revolution, offers several opportunities in the field of health. Indeed, appliances and applications permanently connected to humans and the global digitalization of medical documents produce a vast health data: "Big Health Data". This data is the subject of several projects in the world given the opportunities offered to optimize this area. This paper focuses on quantifying the production of scientific articles about Big Health Data research and the most investigated Big Health Data topics. It also presents a mapping of countries producing articles about this subject. In remote sensing using real time categories, we aimed to quantify articles dealing with "big data architectures", technologies and data sources used. A systematic mapping study was conducted with a set of seven research questions by investigating articles from two digital libraries: Scopus and Springer. The study concern articles published in 2017 and the first half of 2018. The results are illustrated by diagrams answering each question from which a set of recommendations are concluded in this area of research. The study shows that this Data is used the most in studies of oncology. Statistics show that while remote sensing and monitoring is a hot topic, real-time use is not as interesting. It was found that there's a lack in studies interested in big data technologies used in real time remote sensing in the field of health. In conclusion, we recommend more focus on research area treating architecture in remote sensing real time Big Health Data systems combined with geolocation.

## 1. INTRODUCTION

In the age of datamining and real-time medical surveillance, the use of health data takes on a more dynamic dimension, not only informative curative but also predictive and preventive in the long and short term. According to Mr Jonathan Epstein of the National Institute for Child Health and Development, talking about Big Data "All data can tell you something, even if you think it's useless." In this regard, Dr. Michael Rappa from the Institute of Advanced Analysis at North Carolina State University described Big Data as "more opportunistic than scientific"(Coakley et al., 2013). Combined with geolocation, the opportunities will take even more dimensions.

Trying to learn more about the use of Big Data in Health, the gaps and how far technically studies especially in real time remote sensing go, we needed statistical view.

Indeed, we needed first to quantify the production of scientific articles about Big Health Data research in general, most investigated topics, and a mapping of countries producing articles about this subject.

Second, remote sensing is certainly a hot topic but to check if there's a lack in using it in real time, we aimed first to quantify articles using this technology in remote sensing, then dealing with "big data architectures", technologies and data sources used.

In this aim, a systematic mapping study was conducted. First, we define the Big Health Data, the systematic mapping study, data bases and chosen questions. Second, we answer each question with diagrams. Finally, based on this study statistics, we conclude with a set of recommendations.

## 2. BIG HEALTH DATA A SYSTEMATIC MAPPING STUDY

### 2.1 Big Health Data

In 1960, the computerization of the time-consuming and paper-intensive operations of companies led to the beginning of the third technical revolution: the digital revolution. The first being the machine revolution and the second that of the mechanical revolution of the industrial era. (Babinet et al., 2015). In less than sixty years, the digitalization of societies has become complete and global, bringing us into a digital age. Indeed, the standardization of databases and architecture in 1970 allowed applications to feed each other, further encouraging the digitalization of data. The development of microcomputers and local networks in 1980 and the emergence of the Internet and the web have enabled the exchange of messages and documentation between machines and thus the computerization of processes. This movement has not stopped growing since the 2000s and 2010s with the Web 2.0, the widening of the social networks, the appearance of the mobile terminals and the rise of the connected objects (Babinet et al., 2015).

This revolution has led to a need for processing data of different forms that make conventional processing obsolete (Hurwitz et al., 2013).

In the 2000s, web search players were faced with a problem of "scalability" defined by the need to adapt the computing capacity to the rhythms of demand and scalability. Thus, this decade saw the first industrial projects of Big Data but the media appearance of this term to the general public was in 2011 through the report of the American firm McKinsey entitled "Big Data: the new frontier for innovation, competition and productivity "(Huot, 2014).

\* Corresponding author

Big Data is a term consisting of big and data translated according to the Academy of Sciences by "massive data" or data mass (Huot, 2014). According to this name, we understand that Big Data are databases whose volume exceeds the ability to capture, store, manage and analyze typical software. This definition is intentionally subjective. Indeed, with technological advancement, the volume of a database to be considered Big Data will increase as and when. Considering a Big Data database is therefore not fixed by a precise number of bits to be exceeded (Manyika et al., 2011).

The concept of big data is described according to several levels. The four most commonly recognized dimensions represented by 4V are: Volume, Variety, Velocity, and Veracity (Ben Salem, 2015).

The volume represents the amount of data stored. These data are constantly increasing, estimated at 800000 petabytes in the year 2000. It can reach 35 zettabytes in 2020 (Zikopoulos et al., 2012).

The variety describes the diversity of the nature of today's data. Indeed, with the expansion of the use of sensors, telephones and social networks, the data has become complex containing not only traditional data but also semi-structured and unstructured data (web page, streaming, search index...), social networks, forums, e-mail, sensors of active or passive systems ...) making the analysis of traditional way a mission very difficult or impossible (Zikopoulos et al., 2012).

Velocity is the manner and speed at which data is stored and analyzed (Zikopoulos et al., 2012).

The veracity describes the approach that ensures the reliability and quality of data managed and manipulated (Ben Salem, 2015).

To these 4V is added a fifth dimension adopted by some researchers which represents the "Value" parameter. The objective of this 5th V is to allow Big Data to have a meaning and a very precise meaning (Monino et al, 2016).

Technologies that follow our habits, our location, our purchases, our routines, our social interactions and our behaviors are revitalized by mobile phones, downloadable software, monitors and cameras. People are increasingly committed to wearing specialized sensors throughout the day or during exercise or sleep, to provide insight into their physical habits and health (Eagle et al, 2014...). These applications and devices are mostly connected via the Internet to the manufacturer's servers or other analysis or advertising companies and their use is increasing rapidly (Adams et al., 2016).

The vital parameters of the human body are more and more monitored and stored in real time. For example, the San Diego Scripps Clinic's cardiology department has implemented a system that tracks the health status of patients while sitting at home (Agarwal, 2016).

On the other hand, hospital information systems are essential for today's hospitals. They play a crucial role as an information interface for doctors and nurses. A hospital information system includes electronic health records (EHRs), a supplier order entry system, a medical accounting system, an image archiving and communication system (PACS), and so on... Modern hospitals can no longer switch from their information systems that generate large data such as blood test values, electrocardiograms (ECG) and X-rays (X-P), etc. (Sawa, 2014). This situation led to the generation of medical data that was never created before **the Big Health Data**.

These are large data sets collected regularly or automatically and stored electronically. This concept can be reused as multi-purpose data and includes the fusion and connection of existing databases to improve health and health system performance.

These are not data collected for a specific study (Habl et al., 2016).

## 2.2 Systematic mapping study

Systematic mapping studies or scoping studies are designed to give an overview of a research area through classification and counting contributions in relation to the categories of that classification (Petersen et al., 2015).

These studies are similar to systematic reviews, except they employ broader inclusion criteria and are intended to map out topics rather than synthesize study results. They provide a categorical structure for classifying the published research reports and results (Dicheva et al., 2015).

## 2.3 Method

As we needed first to quantify the production of scientific articles about Big Health Data research in general and real time remote sensing in particular, most investigated topics, and a mapping of countries producing articles about this subject, a systematic mapping study was conducted with a set of seven research questions by investigating articles from two digital libraries: Scopus and Springer, setting as studied period 2017 and the first half of 2018:

Q1. What are the hot topics and diseases in the studied period?

Q2. What countries are publishing the most in the studied period?

Q3. At what diseases real time remote sensing is used in the studied period?

Q4. What countries are interested the most in real time remote sensing in the studied period?

Q5. How many publications are citing big data architecture when using real time remote sensing in the studied period?

Q6. What technologies of big data are used in real time remote sensing researches in the studied period?

Q7. What are the data sources when using real time remote sensing in the studied period?

"Big data" and "Health" are the keywords chosen to search in the selected databases, restrictions in research were also used (language, document type, publication year).

Search expressions can be found in Table 1.

Data Base	Advanced or url Search
Springer	facet-language="En"&showAll=false&facet-end-year=2018&facet-start-year=2017&query="big+data"+AND+"health"&facet-content-type="Article"
Scopus	TITLE-ABS-KEY ( "big data" AND "health" ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) ) AND ( LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) )

Table 1. Search in Databases

2365 articles were found after removing articles repeated in the two databases. Also, a restriction was applied as can be seen in table2. Finally, 843 articles were selected to analyze and extract answers based on template in table 3.


Number of articles from search keywords and restrictions	2366
Exclusion criteria: Studies where Big Health Data is not the main subject. Results that are not full text articles (abstracts or editors from proceedings, books...)	
Number of articles after applying exclusion criteria	843

Table 2. Exclusion criteria

Title	Topics	Diseases	Country	Is real time remote sensing used	Does it describe used Big Data	Used Big Data Technology	Data sources
-------	--------	----------	---------	----------------------------------	--------------------------------	--------------------------	--------------

Table 3. Study template

## 2.4 Answers

### 2.4.1 Q1 Answer

Q1. What are the hot topics and diseases in the studied period?  
To answer these question families of Diseases and Topics were created to help full template (table 3). Data extracted based on the 843 selected articles are summarized in table 4 for topics and table 5 for diseases.

Topics	Number of publications
IoT/remote sensing/monitoring (articles using internet of things, remote sensing or distant monitoring)	134
Public health/health management/wellness/personalised medicine (articles about public health, health management of resources and outcomes...)	130
Genomics/ metabolomics	83
Data management	75
Ethics/privacy/security/protocols	63
Medical informatics/precision medicine/Telehealth	57
Epidemiology/Exposures	39
Imagery and radiology	28
web/social media (studies based on Facebook, Twitter...)	23
Drugs and medical product development	11
Brain sciences/neurosciences	11
Toxidocs/toxicity	5
Pharmacovigilance	4
Infection control	1

Table 4. Number of publications per Family of topics

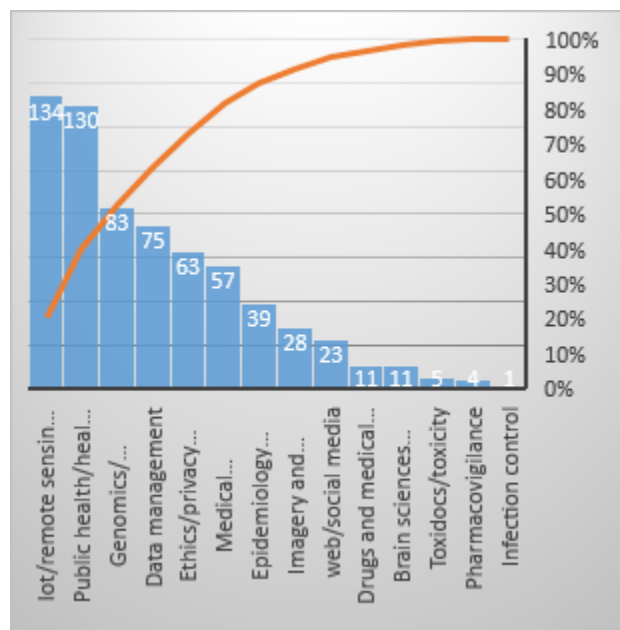


Figure 1. Pareto diagram of publications per Family of topics

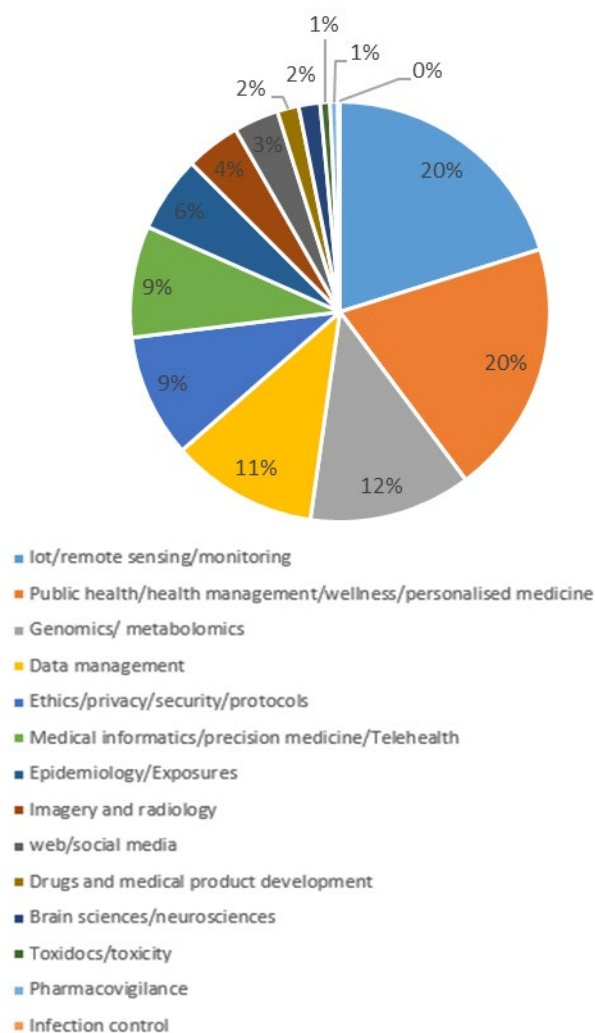


Figure 2. Distribution of publications per Family of topics

Family of disease	Number of publications
Oncology	69
Mental health/Brain sciences/Neurology	56
Cardiology	34
Infectious/Parasite diseases (dengue fever, ebola, malaria...)	30
Pediatrics and Neonatal	14
Diabetes	14
Gastrology	12
Osteology and muscles	10
Hiv	7
Elderly	7
blood pressure diseases	6
Renal diseases	6
Intensive care unit studies	6
Ophthalmology	5
Respiratory illness	5
Dermatology	5
Chronic diseases	4
Blood diseases	4
Rheumatology	4
Others	36

Table 5. Number of publications per Family of Diseases

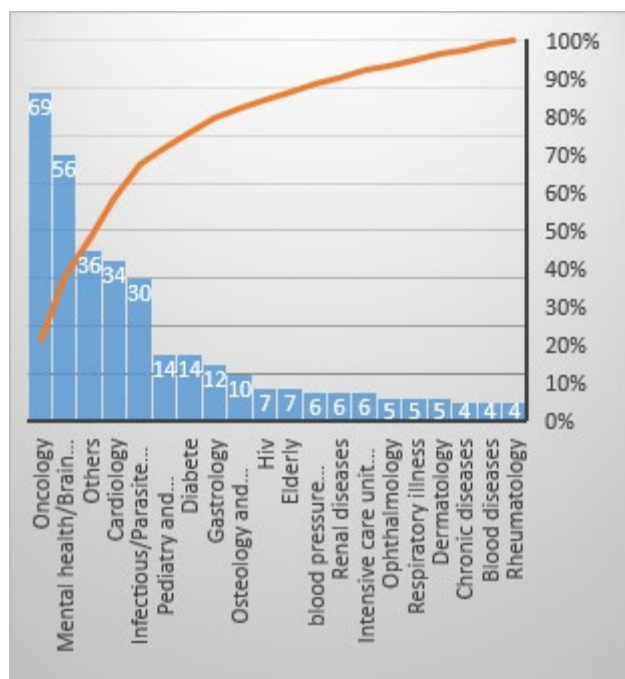


Figure 3. Pareto diagram of publications per Family of Diseases

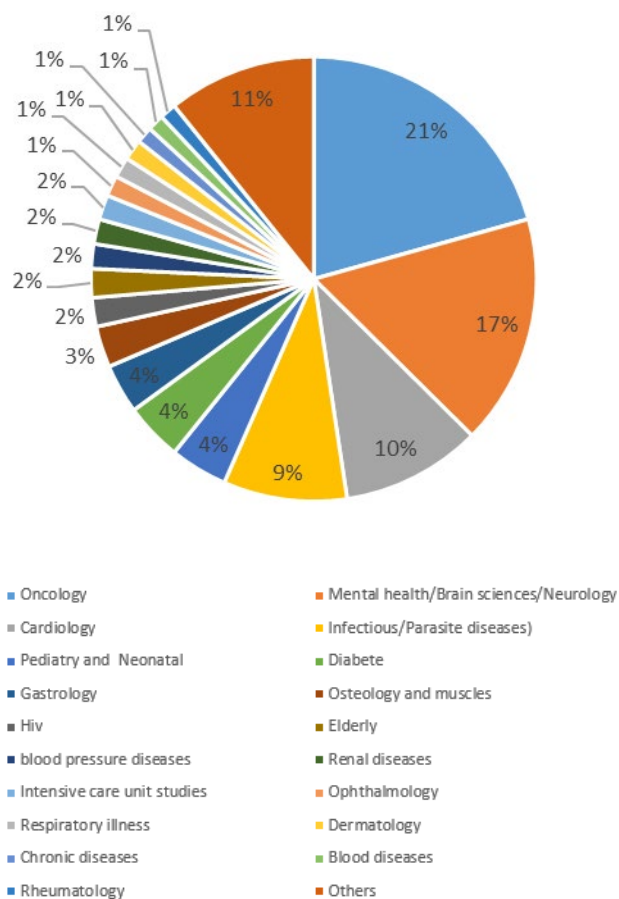


Figure 4. Distribution of publications per Family of Diseases

From figures 1,2,3 and 4, it's seen that Big Health Data studies are interested the most in using internet of things, remote sensing or distant monitoring and health management. For diseases, Big Health Data are used the most in oncology (21%), mental health, neurosciences and brain sciences (17%), cardiology (10%) and in the fourth place it's used in infectious and parasite diseases (9%).

#### 2.4.2 Q2 Answer

Q2. What countries are publishing the most in the studied period?

The top 10 of countries publishing in Big Health Data are Usa, China, United Kingdom, India, Germany, Canada, Australia, Korea, France and ITAY.

Figure 5 presents a mapping of the results, it shows that USA occupy number 1 by publishing 26% of articles, in the second place china by 10% then United Kingdom in the third place by 9 % and India the fourth by 5% of publications.



Figure 5. Mapping of countries number of publications in Big Health Data

### 2.4.3 Q3 Answer

Q3. At what diseases real time remote sensing is used in the studied period?

Based on template (table 3), there are 126 articles dealing with real time remote sensing, which is representing 15% of all articles.

Figure 6 show the percentage of publications in this category per family of diseases.

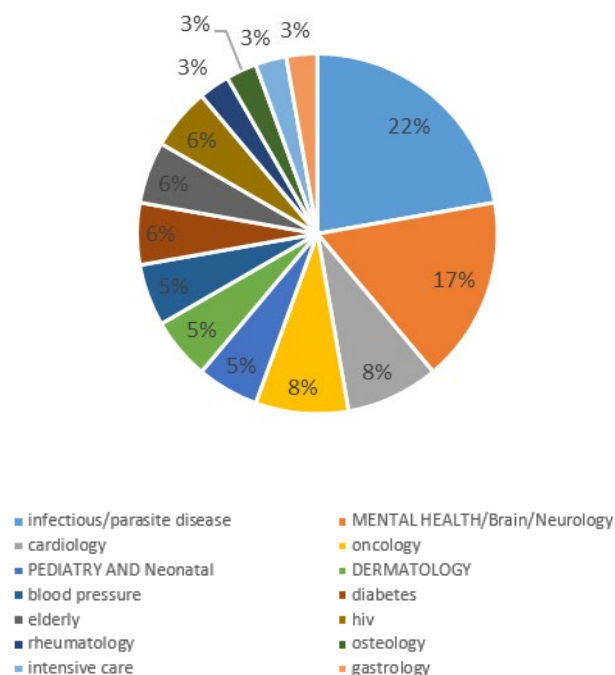


Figure 6. Distribution of real time remote sensing publications per Family of Diseases

In real time remote sensing, researchers are interested the most in monitoring Infectious/Parasite diseases (dengue fever, Ebola, malaria...) (22%), mental health, brain sciences and neurology (17%) then cardiology (8%) and oncology (8%).

### 2.4.4 Q4 Answer

Q4. What countries are interested the most in real time remote sensing in the studied period?

The top 10 of countries publishing in Big Health Data in real time remote sensing is nearly the same as the one of countries publishing in Big Health Data in general.

The top 10 are USA, India, China, United Kingdom, Spain, Korea, Malaysia, Australia, Germany and France.

The difference is that here we find Spain and Malaysia instead of Italy and Canada.

Figure 7 summarizes the results that show that USA is the country publishing the most in both Big Health Data in general and real time remote sensing, but in the second place we find India then China then the United Kingdom instead of being the fourth in all Big Health Data publications.



Figure 7. Mapping of countries number

of publications in Big Health Data real time remote sensing

### 2.4.5 Q5 Answer

Q5. How many publications are citing big data architecture when using real time remote sensing in the studied period?

The results can be seen in Table 6:

Topic	Number of publications
Remote sensing in real time	126
Remote sensing in real time mentioning architecture	51

Table 6. Number of publications in real time remote sensing mentioning big data architecture

Figure 8 shows that only 29% of articles in real time remote sensing are mentioning Big Data architecture.

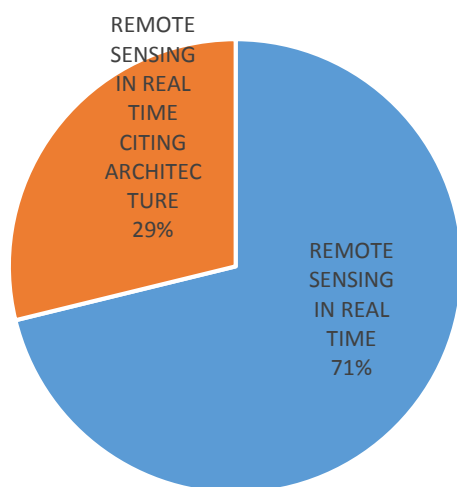


Figure 8. Percentage of publications in real time remote sensing mentioning architecture

#### 2.4.6 Q6 Answer

Q6. What technologies of big data are used in real time remote sensing researches in the studied period?

Technology	Number of publications
cloud	29
hadoop	6
hdfs	6
hbase	5
pig	4
spark	4
mongoDB	2
mapreduce	2
hive	2
storm	1
apache kafka	1
saphana	1
hpcc	1

Table 7: Number of publications per Big Data Technology used in real time remote sensing studies

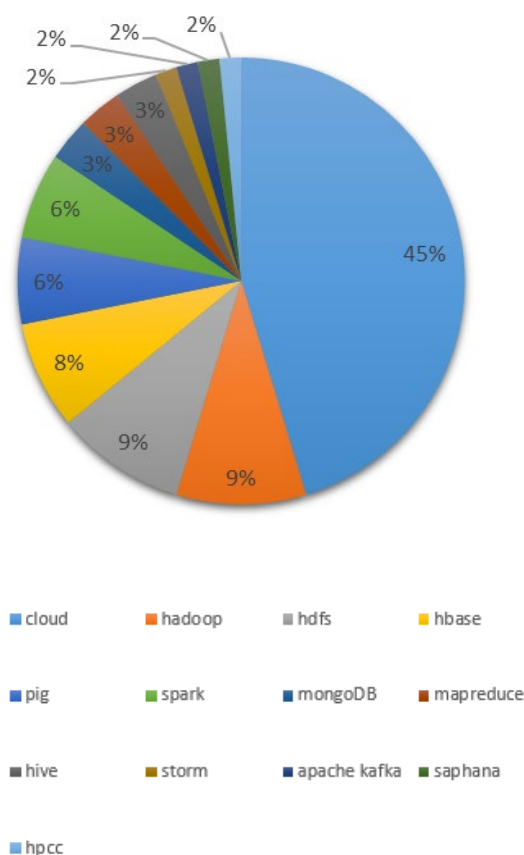


Figure 9. Percentage of publications per Big Data Technology used in real time remote sensing studies

As shown in Figure 9, 45% of articles dealing with the use of Big Health Data in real time remote sensing are only citing the use of Cloud as technology without describing details of architecture.

It's found that Hadoop and Hdfs are the most used Big Data Technologies in this Field.

#### 2.4.7 Q7 Answer

Q7. What are the data sources when using real time remote sensing in the studied period?

It's found that various sources of Big Health Data are used in this type of studies. The advanced in sensors technologies allowed the remote sensing in real time not only of vital signs of the human and Geolocation but also of environmental parameters. Below example of data monitored in real time:

Vital signs: Respiration rate, heart rate, blood pressure diastolic (TAD) and systolic (TAS), ...

Human activity parameters: Sleep periods, calories burned...

Environmental parameters: Humidity, noise, temperature, brightness, the level of Pressure...

#### 2.5 Synthesis

The systematic study was applied on articles published in Spring and Scopus from 2017 to the first half of 2018, 843 articles were selected and used to full template in table 3.

Statistics show that this Data is used the most in studies of oncology, they show also that remote sensing and surveillance, and health management are the hot topics. While remote sensing and surveillance is the hot topic, real time use in remote sensing is not as interesting with a percentage of 15%. From a mapping view, USA is the first country publishing in both Big Health Data in general with 26% and in real time remote sensing category.

The study shows also a lack in studies interested in Big Health Data technologies used in real time remote sensing, in fact from the 15% articles of real time remote sensing only 29% are citing the architecture and from this category 45% are citing the use of cloud with no other technological details.

### 3. CONCLUSION

In conclusion, based on this study, we recommend more focus on research area treating remote sensing real time and more technical studies in Big Health Data architecture.

The study also shows the absence of African countries from the top 10 of articles producers which may open discussions about a gap in Big Health Data in this part of the world.

### REFERENCES

- Adams, S., Purtova, N., Leenes, R., 2016. *Under Observation: The Interplay Between eHealth and Surveillance*. Springer, Berlin, pp. 202.
- Aggarwal, G., 2017. GIS for control of communicable diseases. Geospatial World Forum. Hyderabad India: 23-25 January. <https://geospatialworldforum.org/speaker/SpeakersImages/GIS-for-control-of-communicable-diseases.pdf> (31 October 2018).
- Babinet, G., Vassoyan, R., Asséraf, A., Colligon, H., Delcroix, G., Gerard de Lescazes, F., Graham-Lengrand, S., Munch, J., and Vincent, J., 2015. Big data et objets connectés Faire de la France un champion de la révolution numérique, Report of April 2015, Montaigne Institute, Paris, France.
- Ben Salem, A., 2015. Qualité contextuelle des données : Détection et nettoyage guidés par la sémantique des données, PhD thesis in Computer Science, Paris 13 Sorbonne University, Paris, France.
- Coakley, M.F., Leerkes, M.R., Barnet, J., Gabrielian, A.E., Noble, K., Weber, M.N., and Huyen, Y., 2013. Unlocking the Power of Big Data at the National Institutes of Health <https://doi.org/10.1089/big.2013.0012>.
- Dicheva, D., Dichev, C., Agre, G., Angelova, G., 2015. Gamification in Education: A Systematic Mapping Study [https://www.researchgate.net/publication/270273830\\_Gamification\\_in\\_Education\\_A\\_Systematic\\_Mapping\\_Study](https://www.researchgate.net/publication/270273830_Gamification_in_Education_A_Systematic_Mapping_Study) (31 October 2018).
- Eagle, N., Greene, K., 2014. *Reality Mining: Using Big Data to Engineer a Better World*. MIT Press, London, pp. 11.
- Habl, C., Renner, A-T., Bobek, J., Laschkolnig, A., 2016. Study on Big Data in Public Health, Telemedicine and Healthcare, Final Report of October 2016, Gesundheit Österreich GmbH, Vienna, Austria.
- Huot, C., 2014. Les Big Data Si nous en parlions ? In *Big Data, Nouvelles partitions de l'information*. De Boeck, Louvain-la-Neuve, pp. 13- 10.
- Hurwitz, J., Nugent, A., Halper, F., Kaufman, M., 2013. *Getting Started with Big Data. Big data for dummies*. John Wiley & Sons, Inc, New Jersey, pp. 10–21.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity, Report of June 2011, McKinsey Global Institute, New York, USA.
- Monino, J., Sedkaoui, S., 2016. *Big Data, Open Data and Data Development*. John Wiley & Sons, New York, pp. 102.
- Petersen, K., Vakkalanka, S., Kuzniarz, L., 2015. Guidelines for conducting systematic mapping studies in software engineering: An update <http://dx.doi.org/10.1016/j.infsof.2015.03.007>.
- Sawa, T., 2014. Leading Advances in the Utilization of Big Data in the Healthcare Industry; Intel Health & Life Sciences <https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-healthcare-tokyo-paper.pdf> (31 October 2018).
- Zikopoulos, P., Eaton, C., Deroos, D., Deutsch, T., Lapis, G., 2012. *Big Data: From the Business Perspective. Understanding Big Data*. McGraw-Hill, New York, pp. 5-7.