Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 639021, 9 pages
http://dx.doi.org/10.1155/2015/639021

Review Article

Toward a Literature-Driven Definition of Big Data in Healthcare

Emilie Baro, Samuel Degoul, Régis Beuscart, and Emmanuel Chazard

Department of Public Health, EA 2694, University of Lille, 1 Place de Verdun, 59045 Lille Cedex, France

Correspondence should be addressed to Emilie Baro; emilie.baro@gmail.com

Received 13 November 2014; Accepted 4 February 2015

Academic Editor: Shahram Shirani

Objective. The aim of this study was to provide a definition of big data in healthcare. *Methods.* A systematic search of PubMed literature published until May 9, 2014, was conducted. We noted the number of statistical individuals (n) and the number of variables (p) for all papers describing a dataset. These papers were classified into fields of study. Characteristics attributed to big data by authors were also considered. Based on this analysis, a definition of big data was proposed. *Results.* A total of 196 papers were included. Big data can be defined as datasets with $\text{Log}(n * p) \geq 7$. Properties of big data are its great variety and high velocity. Big data raises challenges on veracity, on all aspects of the workflow, on extracting meaningful information, and on sharing information. Big data requires new computational methods that optimize data management. Related concepts are data reuse, false knowledge discovery, and privacy issues. *Conclusion.* Big data is defined by volume. Big data should not be confused with data reuse: data can be big without being reused for another purpose, for example, in omics. Inversely, data can be reused without being necessarily big, for example, secondary use of Electronic Medical Records (EMR) data.

1. Introduction

The 21st century is an era of big data involving all aspects of human life, including biology and medicine [1]. With the advance in genomics, proteomics, metabolomics, and other types of omics technologies during the past decades, a tremendous amount of data related to molecular biology has been produced [2]. In addition, the transition from paper medical records to EHR systems has led to an exponential growth of data [3]. As a result, big data provides a wonderful opportunity for physicians, epidemiologists, and health policy experts to make data-driven decisions that will ultimately improve patient care [3]. As Margolis stated, "Big data are not only a new reality for the biomedical scientist, but an imperative that must be understood and used effectively in the quest for new knowledge" [4].

To date, however, the term "big data" does not have a proper definition in the MeSH (Medical Subject Headings) database yet. A precise, well-formed, and unambiguous definition is a requirement for a shared understanding of the term big data. The objective of this work is to provide a definition of big data in healthcare through a review of the literature.

2. Material and Methods

2.1. Search Strategy. For this literature review, we conducted a systematic search of the PubMed database for all papers published until May 9, 2014, using the keywords "big data." To be fully inclusive, we did not define a start date. We used the following PubMed query:

(a) (big data[Title/Abstract]) AND ("1900/01/01"[Date - Publication]: "2014/05/09"[Date - Publication]).

Titles and abstracts were reviewed by a human for eligibility. Papers were excluded if they were not directly related to healthcare or if big data was not found to be the topic of the paper.

We then attempted to retrieve the full-text papers. We used online search facilities (the Free PMC database, Google, and Google Scholar), resources, and services of the Lille University library and tried to directly contact the first or corresponding author. Full-text papers were then read.

Each of the remaining papers was included in the analysis and classified either as a paper describing a dataset, a dissertation, or a review of the literature.

2.2. Data Collection Process. For each paper, we collected the following information: title, year of publication, journal title, specialty area, type of paper (paper using a dataset, dissertation, and literature review), the field of study, and characteristics given by authors to big data and to data reuse. In case the paper dealt with a dataset, we also collected the number of statistical individuals (n) and the number of variables (p). It should be noted that the number of statistical individuals n is not necessarily physical persons but can also be, for example, gene sequences. The number of variables p could be, for example, the number of physicochemical properties used to classify amino acids [5], the performance metrics adopted to evaluate model performance [6], or the number of features of medical claims. In this last case, the number of individuals n is represented by the number of records of medical claims [7].

2.3. Analysis and Classification. Statistical analyses were performed with R statistical computing software [8]. In this paper, the notation "Log" denotes the decimal (or common, or decadic) logarithm, and the notation "CI_{95}" denotes 95% confidence intervals. CI_{95} of binary variables were computed using the binomial law.

2.3.1. Time Evolution of Publication about Big Data in Healthcare. To analyze the evolution of publication in healthcare, we draw a graph showing the annual publication of papers included in our review and a graph showing the annual publication of papers which were describing a dataset. We also noted the number of journals which published papers about big data in healthcare per year.

2.3.2. Time Evolution of the Size of Big Data in Healthcare. In order to see the evolution of what authors refer to as "big data," from papers describing a dataset, we plotted the decimal logarithm of the product of the number of statistical individuals (n) and the number of variables (p), $\text{Log}(n * p)$, as a function of the year.

2.3.3. Number of Individuals and Variables in Each Field of Study. The numbers n and p were analyzed with respect to the field of study. To this end, the probability density functions of $\text{Log}(n)$, $\text{Log}(p)$, and $\text{Log}(n * p)$ were plotted with respect to fields of study. Finally, $\text{Log}(p)$ as a function of $\text{Log}(n)$ was plotted with respect to fields of study.

2.4. Characteristics of Big Data. Characteristics attributed to big data by the authors in free text were noted as reading all the papers included in the analysis and were then sorted out by categories.

2.5. Proposal of a Definition of Big Data. We then gathered to propose a definition of big data in healthcare. A difference was made between definition, properties, and related concepts. A dataset that matches the definition qualifies as "big data," and thus has the properties that are proposed. Conversely, a dataset that has some or all of the listed properties does not necessarily qualify as "big data." Finally, related concepts refer to properties that are not systematically related to big data.

We attempted to bring out a threshold of the volume of big data on the basis of findings from this literature review. The threshold resulted from a discussion between the authors of this paper, taking into account sizes of actual datasets, but also properties that are attributed to big data by the authors of the papers included in this literature review.

Table 1: Number of papers by field of study among the 48 papers describing a dataset.

Field of study	Number of papers
Omics	
Genomics	18
Metabolomics	1
Proteomics	4
Medical specialties	
Endocrinology	2
Imaging	3
Immunology	1
Infectiology	1
Neurology	8
Pharmacovigilance	1
Public health	
Bioinformatics	3
EHR*	1
Epidemiology	2
Public health	3

*EHR: Electronic Health Records.

3. Results

3.1. Search Strategy. The search query yielded 330 papers. After reading titles and abstracts, 94 papers were excluded. A total of 236 papers were included for full-text review. Eighteen papers were unavailable. The full-texts of the remaining 218 papers were read. After applying the exclusion criteria, 22 papers were excluded, leaving 196 papers. Papers were excluded due to the following reasons: papers not directly related to healthcare (18 papers) and papers in which big data was not the topic of the paper (4 papers). Of the 196 papers left for inclusion, there were 48 papers describing a dataset, 121 dissertations, and 27 reviews of the literature. Figure 1 shows a detailed description of the search strategy and results.

3.2. Data Collection Process. The number of papers by field of study among the 48 papers describing a dataset is listed in Table 1.

Among the 48 papers describing a dataset, three main categories of studies were identified: omics, medical specialties, and public health. The term "omics" refers to biology fields of study ending in -omics, such as genomics, metabolomics, or proteomics. The main area represented is omics: 23 papers (48%, $CI_{95} = [33; 63]$). It is followed by medical specialties (endocrinology, infectology, immunology, neurology, and

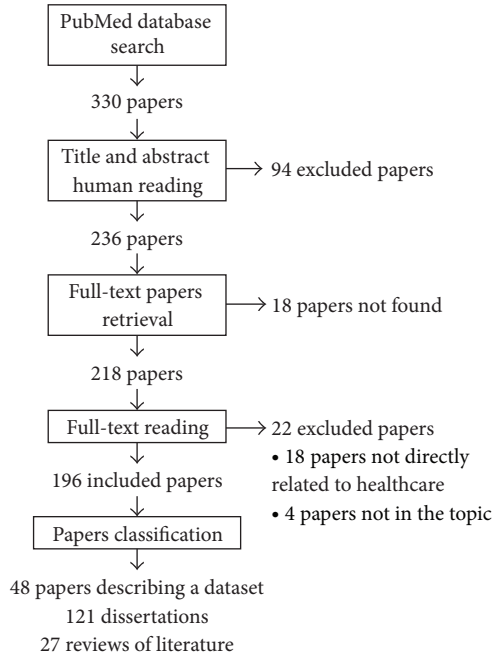

Figure 1: Flowchart of the literature review.

imaging): 15 papers (31%, CI_{95} = [19; 46]) and public health (bioinformatics, Electronic Health Records (EHR), epidemiology, pharmacovigilance, and public health): 10 papers (21%, CI_{95} = [10; 35]).

3.3. Analysis and Classification

3.3.1. Time Evolution of Publication about Big Data in Healthcare. Figure 2 shows the evolution of the publication of papers about big data in healthcare from 2003 to 2013. Annual publication of papers about big data in healthcare increased from 1 in 2003 to 79 in 2013. In the same way, an increase in the annual publication of papers describing a dataset can be observed (Figure 3). The 196 papers included in our review were published in 134 different journals. Among these journals, one journal published papers about big data in healthcare in 2008. There were 68 in 2013.

3.3.2. Time Evolution of the Size of Big Data in Healthcare. Figure 4 illustrates the decimal logarithm of the number of statistical individuals multiplied by the number of variables ($\text{Log}(n * p)$) for each year of publication of the papers that describe a dataset. We observe a nonsignificant increase of 0.43 per year (P value = 0.34).

3.3.3. Number of Individuals and Variables in Each Field of Study. Figures 5, 6, and 7 represent the probability density function of $\text{Log}(n)$, $\text{Log}(p)$, and $\text{Log}(n * p)$, respectively, for omics, medical specialties, public health, and all papers. It can be pointed out that $\text{Log}(n * p)$ is inferior to 7 in 23 studies out of 48 (48%, CI_{95} = [33; 63]).

Figure 8 shows $\text{Log}(p)$ as a function of $\text{Log}(n)$ for omics, medical specialties, and public health. This figure suggests

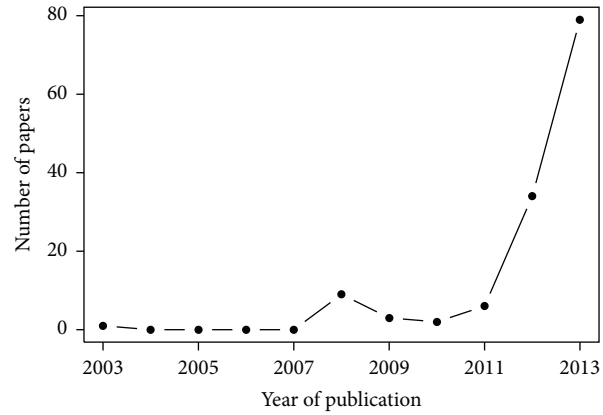

Figure 2: Number of papers about big data in healthcare published per year (full years only).

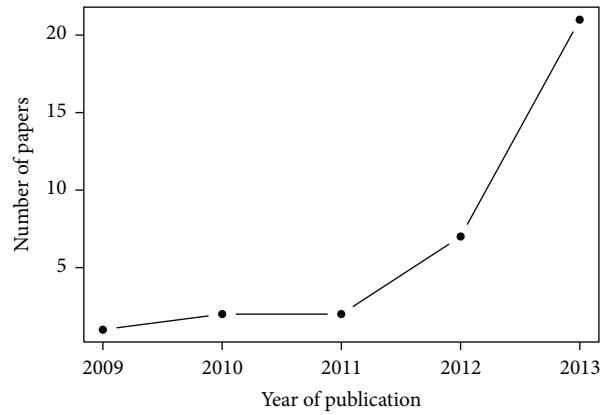

Figure 3: Number of papers about big data in healthcare describing a dataset per year (full years only).

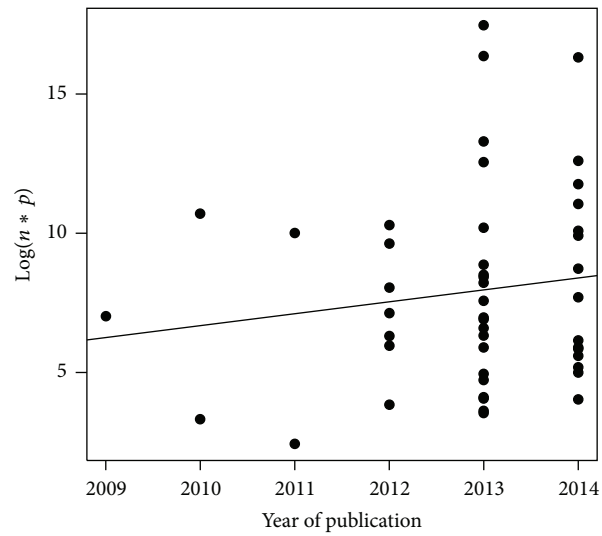

Figure 4: $\text{Log}(n * p)$ per year of publication. The continuous line represents the linear regression (P = 0.34).

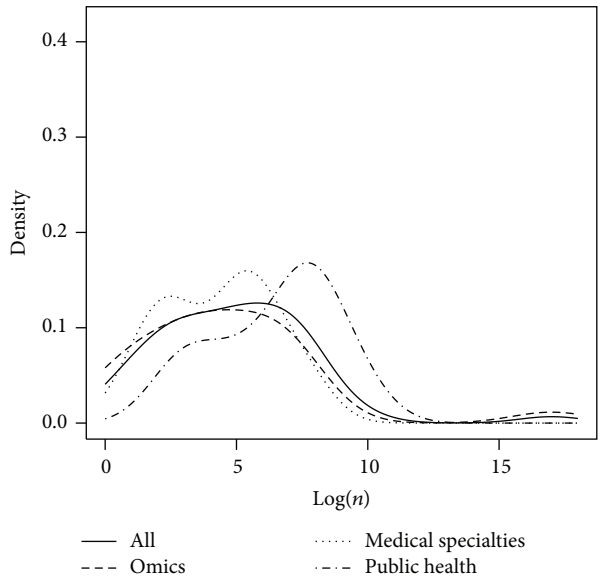

Figure 5: Representation of the probability density function of Log(n) for omics, medical specialties, public health, and all fields together.

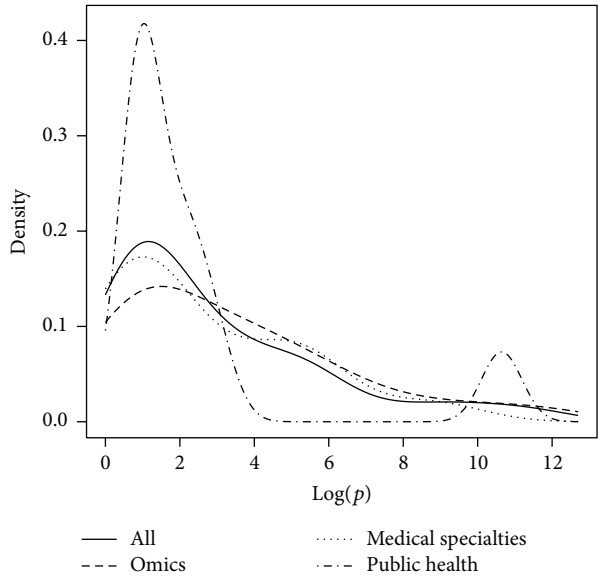

Figure 6: Representation of the probability density function of Log(p) for omics, medical specialties, public health, and all fields together.

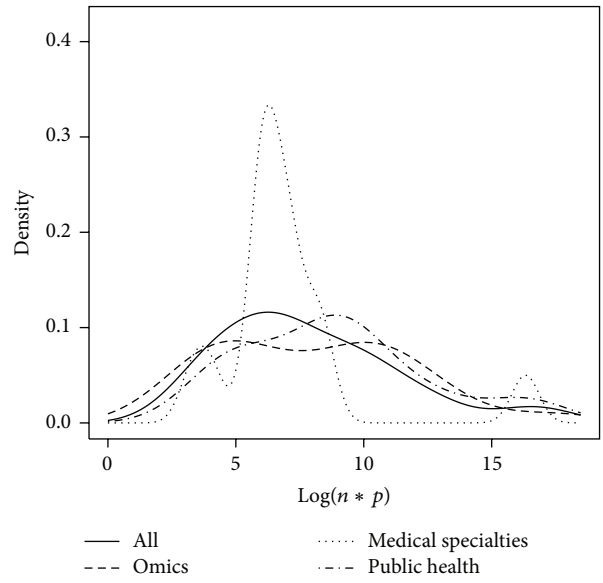

Figure 7: Representation of the probability density function of Log($n * p$) for omics, medical specialties, public health, and all fields together.

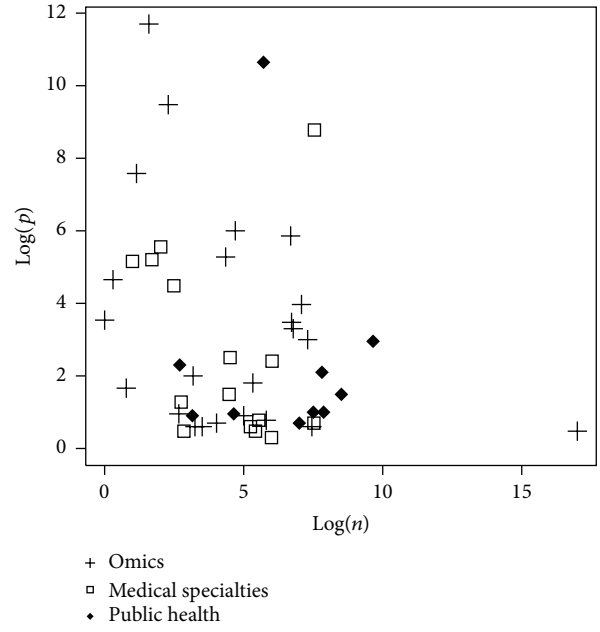

Figure 8: Log(p) as a function of Log(n) for omics, medical specialties, and public health. Each pictogram stands for one paper.

the following differences between omics, medical specialties, and public health categories:

(i) big data in omics concern massive data collected on a limited number of individuals: small n, high p;

(ii) public health studies concern an important number of individuals and a low number of variables: high n, small p;

(iii) medical specialties are characterized by an important number of individuals and variables: high n, high p.

3.4. Characteristics of Big Data. The main characteristic about big data found in the papers is its massive size and complexity [7, 9–17]. Big data concern "not only the sheer scale and breadth of the new data sets but also their increasing complexity" [15]. Widely used notions to describe the complexity of big data are the three "Vs": volume, variety, and velocity [7, 18–25]. "Big Data is a term used to describe information assemblages that make conventional data, or database, processing problematic due to any combination

of their size (volume), frequency of update (velocity), or diversity (variety)" [18]. Veracity is a fourth "V" sometimes added to describe big data challenge [17, 23, 26–28]. Some authors mention a fifth "V": valorization [26, 29].

3.4.1. Volume. Volume is the main characteristic mentioned by authors [7, 12, 16, 21, 23, 26, 30, 31]. "These correspond to the well-accepted notions of volume (breadth and/or depth) (…) recognized as the hallmarks of big data" [21]. "For volume, this translates today into terabytes (10^{12} bytes), petabytes (10^{15} bytes) or exabytes (10^{18} bytes)" [7]. "Volume - much greater amounts of rapidly multiplying data than were ever previously available" [25]. Some authors mention a big data threshold without clearly defining it [7, 32]: "How big is 'Big'? (…) size is a relative term when it comes to data" [32]. "Those data are unquestionably 'big' (order 10^{17})" [21]. Data sets used "in epidemiology (…) in fact barely pass the 'big data' threshold" [7].

3.4.2. Variety. Variety is another important characteristic of big data [7, 25, 26, 30, 31, 33–35]. Indeed, big data comes from various sources [23, 36]. Variety translates into "aggregation of widely disparate sources of data or mash-ups of data derived from independent sources" [7]. Unstructured data, for example, free text data [7, 12, 37] and images [32, 38–40], are particularly a big challenge. In healthcare, "data take many forms including numbers, text, coded data, graphics, images, physiological measures (signals), and sound. Healthcare professionals rely on all their senses, including smell, to collect assessment data from individuals" [12]. In this area, "unstructured data is expected to exponentially outpace structured data" [34]. "Electronic Medical Records (EMR) generate massive data sets, offering the challenge of how to convert largely unstructured by-products of healthcare delivery into useful assets for patients' insight" [41]. Big data "can deviate from traditional structured data (organized in rows and columns) and can be represented as semi-structured data such as XML, or unstructured data including flat files which are not compliant with traditional database methods" [33]. These data are "unstructured for analysis using conventional relational database techniques" [31].

Moreover, big data can be "volatile, that is, changing, and available only for a limited amount of time" [23].

3.4.3. Velocity. Accelerated increase of data is another attribute of big data [7, 21, 23, 25, 26, 31, 42]. It is "data at or near real-time" [25]. "Velocity refers to the enormous frequency with which today's data is generated, delivered, and processed" [31].

3.4.4. Challenge on Veracity. Veracity comes next: big data can be difficult to validate [17, 26–28]. "Big data must be interpreted with caution, and in context, if it is to be clinically useful" [27]. It has a low veracity. Big data can never "be 100% accurate" [28].

3.4.5. Challenges on All Aspects of the Workflow. Big data raises challenges on all aspects of the workflow: from amassing [32], capturing [7, 37, 43–45], collecting [20, 46], storing [7, 20, 32, 43, 44, 47–53], data management [20, 43, 45, 54, 55], processing [9, 12, 19, 26, 47, 48, 51, 52, 56, 57], and analyzing [7, 20, 31–33, 39, 43–45, 49–55, 58–60], to peer-reviewed publications of results [45]. Big data "creates difficulties in data capture, storage, cleaning, analytics, visualization and sharing" [43]. Big data is also difficult to valorize [26, 29]: big data "is not merely large in volume; it also moves rapidly, is difficult to validate and valorize" [26].

3.4.6. Challenges on Statistical and Computational Methods. Finding new statistical and computational methods is another challenge raised by big data [33, 43, 50, 51, 59, 61, 62]. Big data requires "a change of perspective, infrastructure, and methods for data collection and analyses" [62]. Visualization methods that allow us to understand the data need to be created [32, 43, 44, 57]. To make sense of big data, "the further creation of new tools and services for data discovery, integration, analysis, and visualization" [32] will be required.

3.4.7. Challenges on Extracting Meaningful Information. Several authors emphasize the fact that it is necessary to derive useful information of these data [30, 44, 63, 64] and raise the question of how the data could be meaningfully interpreted: big data creates "challenges around how to meaningfully interpret the data - much of it not described using consistent standards or metadata - into information and recommendations while eliminating noise and erroneous data" [19].

3.4.8. Challenges on Facilitating Information Access and Sharing. Many authors highlight the necessity of identifying ways to facilitate information access and sharing [7, 15, 30, 34, 43–46, 49, 50, 53, 62, 63, 65–67]. It is necessary to promote "collaboration among scientists" [46]. Data must be made more readily available from more open sources to better compare data.

3.4.9. Not Enough Human Experts. Some authors mention the fact that the number of available human experts who have both clinical and analytic knowledge is not sufficient yet [30, 68]: "the role needs some sort of hybrid person that has clinical knowledge and analytic knowledge. We are experiencing a drought in terms of analytic experience. We don't have enough of those people in place yet" [30].

3.4.10. Data Reuse. Some authors mention the fact that big data can be data that are commonly collected without an immediate use: "Massive amounts of data are commonly collected without an immediate business case, but simply because it is affordable. This data, so it is hoped, will later answer questions, most of which yet have to arise" [20]. They put into light the fact that big data are often a secondary use of data, which we can call data reuse [14, 20, 21, 41, 65, 69–72].

3.4.11. False Knowledge Discovery. Some authors highlight the fact that deriving knowledge from big data can lead to false results and to conclusions that are wrong [73–75]: "Exploratory results emerging from Big Data are no less likely

to be false" [75]. We cannot extract knowledge from big data without knowing the context in which data sets were collected: "big size is not enough for credible epidemiology" [74].

3.4.12. Privacy Issues. One concern mentioned by several authors is privacy issues: "the increasing ease with which data may be used and reused has increased concerns about privacy and informed consent" [76]. The ability "to protect individual privacy in the era of big data has become limited" [39]. Even if large databases use pseudonymised personal confidential data that have been anonymised, they retain a residual risk of reidentification. Indeed, the identity of individuals can be determined by manipulating databases through data linkage techniques [28, 39, 66, 77]. The data torrent poses ethical challenges [15]. "The widespread implementation of EHRs and the need to share data to measure quality and manage accountable care organizations (ACOs) brings to light all of the privacy issues surrounding sharing patient data" [66]. "The ability to derive DNA-based information from non-DNA-based sources generalizes the issue of data de-identification beyond the area of genotypic data privacy and has thus potentially important consequences for privacy rules in scientific research" [39].

3.5. Proposal of a Definition of Big Data. A definition of big data was established on the basis of findings from the literature review. We consider that big data should exclusively be defined by volume, and we propose that a dataset could be qualified as "big dataset" only if $\text{Log}(n * p)$ is superior or equal to 7.

Properties of big data can be listed as follows:

(i) great variety,
(ii) high velocity,
(iii) challenge on veracity,
(iv) challenge on all aspects of the workflow,
(v) challenge on computational methods,
(vi) challenge on extracting meaningful information,
(vii) challenge on sharing data,
(viii) challenge on finding human experts.

Related concepts of big data are as follows:

(i) data reuse,
(ii) false knowledge discovery,
(iii) privacy issues.

The definition of big data is summed up in Table 2.

Table 2: Definition of big data in healthcare.

Definition	Volume: $\text{Log}(n * p) \geq 7$
Properties	Great variety
	High velocity
	Challenge on veracity
	Challenge on all aspects of the workflow
	Challenge on computational methods
	Challenge on extracting meaningful information
	Challenge on sharing data
	Challenge on finding human experts
Related concepts	Data reuse
	False knowledge discovery
	Privacy issues

4. Discussion

In this work, through a detailed literature review, we tried to provide a current and quantitative definition of big data. We performed a literature review of 196 papers published until May 2014. Finally, we proposed a definition of big data in healthcare.

This systematic search should ensure that we accumulate a relatively complete census of relevant literature of big data in healthcare. However, we may have missed papers that do use big data in the research but were not included in our query because the term was not mentioned in the abstract or keywords of the paper. Those papers could be less and less frequent in the future.

Nevertheless, as there is no definition of big data, the literature can itself be wrong. It is a limitation of this inductive approach: we use observations to build a definition. The problem of defining a threshold illustrates this difficulty: the threshold of 10^7 may appear in disagreement with the results of Figure 7. This definition of big data is simply the result of a discussion between the authors of this literature review. It has been decided based on the results of the number of individuals and of variables found in the studies describing a dataset, but it has also taken into account the characteristics of big data mentioned by the authors of all the papers included in this literature review. Thus, for example, we can consider that the problems related to computational methods do not exist for $\text{Log}(n * p)$ inferior to 7, even when the analysis is performed with a simple spreadsheet instead of statistical software calling for high computational capacities. However, this proposal suggests that half of the studies describing a dataset in this literature review wrongly call their dataset big data. As everyone talks about the challenges of computing and data processing, considering what we know today in practice about software and computers, it would have been difficult to admit a threshold of $\text{Log}(n * p)$ superior or equal to 6 (although such a threshold already excludes 35% of the studies of our review), because we know that, nowadays, such size of data is easy to deal with.

It should also be pointed out that there is an undeniable current trend of big data, which leads to the fact that the term "big data" is now used to qualify datasets that, in the past, would not have been called this way. Moreover, we can consider that the size of datasets that qualify as big data may keep on increasing due to the main property of big data, which is the challenge on data processing and the fact that computational infrastructure that is required to process these large-scale datasets may progress with time.

Data reuse has been defined as a related concept of big data because we think that there might be some confusion between these two terms: data reuse is the fact of using for decisional purposes data that were collected routinely for transactional purposes, whereas big data is related to the size of the data collection. Indeed, data can be big without being reused for another purpose: this is the case of omics, for example. Inversely, data can be reused without being necessarily big, such as secondary use of data from Electronic Medical Records (EMR).

Big data presents many opportunities for translational studies, and informatics will be the key for successful translational research [78]. As Shah stated, "translational informatics is ready to revolutionize human health and healthcare using large-scale measurements on individuals. Data-centric approaches that compute on massive amounts of data to discover patterns and to make clinically relevant predictions will gain adoption" [79]. Cloud computing could be an enabling tool to facilitate translational bioinformatics research [67].

Informatics is needed to fully harness the potential of health data and new tools are emerging to translate health data into knowledge for improved healthcare.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] Z. Zhang, "Big data and clinical research: focusing on the area of critical care medicine in mainland China," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 5, pp. 426–429, 2014.

[2] S. Li, L. Kang, and X.-M. Zhao, "A survey on evolutionary algorithm based hybrid intelligence in bioinformatics," *BioMed Research International*, vol. 2014, Article ID 362738, 8 pages, 2014.

[3] D. I. Sessler, "Big Data—and its contributions to peri-operative medicine," *Anaesthesia*, vol. 69, no. 2, pp. 100–105, 2014.

[4] R. Margolis, L. Derr, M. Dunn et al., "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data," *Journal of the American Medical Informatics Association*, vol. 21, no. 6, pp. 957–958, 2014.

[5] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, and Z. Lin, "An approach for identifying cytokines based on a novel ensemble classifier," *BioMed Research International*, vol. 2013, Article ID 686090, 11 pages, 2013.

[6] L. Zhao, L. Wong, L. Lu, S. C. H. Hoi, and J. Li, "B-cell epitope prediction through a graph model," *BMC Bioinformatics*, vol. 13, supplement 17, p. S20, 2012.

[7] M. L. Berger and V. Doban, "Big data, advanced analytics and the future of comparative effectiveness research," *Journal of Comparative Effectiveness Research*, vol. 3, no. 2, pp. 167–176, 2014.

[8] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, http://www.r-project.org/.

[9] W. J. Mallon, "Big data," *Journal of Shoulder and Elbow Surgery*, vol. 22, no. 9, article 1153, 2013.

[10] R. S. Salcido, "Big data and disruptive innovation in wound care," *Advances in Skin and Wound Care*, vol. 26, no. 8, article 344, 2013.

[11] T. Ketchersid, "Big data in nephrology: friend or foe?" *Blood Purification*, vol. 36, no. 3-4, pp. 160–164, 2014.

[12] E. J. S. Hovenga and H. Grain, "Health data and data governance," *Studies in Health Technology and Informatics*, vol. 193, pp. 67–92, 2013.

[13] H. Müller, A. Hanbury, and N. Al Shorbaji, "Health information search to deal with the exploding amount of health information produced," *Methods of Information in Medicine*, vol. 51, no. 6, pp. 516–518, 2012.

[14] D. J. Porche, "Men's health big data," *American Journal of Men's Health*, vol. 8, no. 3, p. 189, 2014.

[15] W. Callebaut, "Scientific perspectivism: a philosopher of science's response to the challenge of big data biology," *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, no. 1, pp. 69–80, 2012.

[16] J. Fan and H. Liu, "Statistical analysis of big data on pharmacogenomics," *Advanced Drug Delivery Reviews*, vol. 65, no. 7, pp. 987–1000, 2013.

[17] O.-S. Lupşe, M. Crisan-Vida, L. Stoicu-Tivadar, and E. Bernard, "Supporting diagnosis and treatment in medical care based on big data processing," *Studies in Health Technology and Informatics*, vol. 197, pp. 65–69, 2014.

[18] S. I. Hay, D. B. George, C. L. Moyes, and J. S. Brownstein, "Big data opportunities for global infectious disease surveillance," *PLoS Medicine*, vol. 10, no. 4, Article ID e1001413, 2013.

[19] B. Hamilton, "Impacts of big data. Potential is huge, so are challenges," *Health Management Technology*, vol. 34, no. 8, pp. 12–13, 2013.

[20] A. Markowetz, K. Błaszkiewicz, C. Montag, C. Switala, and T. E. Schlaepfer, "Psycho-informatics: big data shaping modern psychometrics," *Medical Hypotheses*, vol. 82, no. 4, pp. 405–411, 2014.

[21] C. G. Chute, M. Ullman-Cullere, G. M. Wood, S. M. Lin, M. He, and J. Pathak, "Some experiences and opportunities for big data in translational research," *Genetics in Medicine*, vol. 15, no. 10, pp. 802–809, 2013.

[22] R. R. Kao, D. T. Haydon, S. J. Lycett, and P. R. Murcia, "Supersize me: how whole-genome sequencing and big data are transforming epidemiology," *Trends in Microbiology*, vol. 22, no. 5, pp. 282–291, 2014.

[23] K. Sedig and O. Ola, "The challenge of big data in public health: an opportunity for visual analytics," *Online Journal of Public Health Informatics*, vol. 5, no. 3, article 223, 2014.

[24] E. Gardner, "The HIT approach to big data," *Health data management*, vol. 21, no. 3, pp. 34–38, 2013.

[25] K. D. Moore, K. Eyestone, and D. C. Coddington, "The big deal about big data," *Healthcare Financial Management*, vol. 67, no. 8, pp. 60–68, 2013.

[26] T. Dereli, Y. Coşkun, E. Kolker, Ö. Güner, M. Ağirbaşli, and V. Özdemir, "Big data and ethics review for health systems research in LMICs: understanding risk, uncertainty and ignorance-and catching the black swans?" *American Journal of Bioethics*, vol. 14, no. 2, pp. 48–50, 2014.

[27] R. S. Litman, "Complications of laryngeal masks in children: big data comes to pediatric anesthesia," *Anesthesiology*, vol. 119, no. 6, pp. 1239–1240, 2013.

[28] J. C. Ward, "Oncology reimbursement in the era of personalized medicine and big data," *Journal of Oncology Practice*, vol. 10, no. 2, pp. 83–86, 2014.

[29] V. Özdemir, K. F. Badr, E. S. Dove et al., "Crowd-funded micro-grants for genomics and 'big data': an actionable idea connecting small (Artisan) science, infrastructure science, and citizen philanthropy," *OMICS*, vol. 17, no. 4, pp. 161–172, 2013.

[30] AHA, "Harnessing big data: how to achieve value," *Hospitals & Health Networks*, vol. 88, no. 2, pp. 61–71, 2014.

[31] K. Jee and G.-H. Kim, "Potentiality of big data in the medical sector: focus on how to reshape the healthcare system," *Healthcare Informatics Research*, vol. 19, no. 2, pp. 79–85, 2013.

[32] J. D. van Horn and A. W. Toga, "Human neuroimaging as a 'Big Data' science," *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 323–331, 2014.

[33] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781, 2013.

[34] "Buyer's brief: cognitive computing in the age of big data," *Healthcare Financial Management*, vol. 68, no. 4, pp. 35–36, 2014.

[35] T. H. Davenport and D. J. Patil, "Data scientist: the sexiest job of the 21st century," *Harvard Business Review*, vol. 90, no. 10, pp. 70–128, 2012.

[36] M. J. Khoury, T. K. Lam, J. P. A. Ioannidis et al., "Transforming epidemiology for 21st century medicine and public health," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 22, no. 4, pp. 508–516, 2013.

[37] S. Bonney, "HIM's role in managing big data: turning data collected by an EHR into information," *Journal of American Health Information Management Association*, vol. 84, no. 9, pp. 62–64, 2013.

[38] C. P. Jayapandian, C.-H. Chen, A. Bozorgi, S. D. Lhatoo, G.-Q. Zhang, and S. S. Sahoo, "Cloudwave: distributed processing of 'big data' from electrophysiological recordings for epilepsy clinical research using hadoop," *AMIA Annual Symposium Proceedings*, vol. 2013, pp. 691–700, 2013.

[39] E. E. Schadt, "The changing privacy landscape in the era of big data," *Molecular Systems Biology*, vol. 8, article 612, 2012.

[40] A. Aji, F. Wang, and J. H. Saltz, "Towards building a high performance spatial query system for large scale medical imaging data," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12)*, pp. 309–318, 2012.

[41] G. O. Matheson, M. Klügl, L. Engebretsen et al., "Prevention and management of noncommunicable disease: the IOC consensus statement, lausanne 2013," *Clinical Journal of Sport Medicine*, vol. 23, no. 6, pp. 419–429, 2013.

[42] F. M. Afendi, N. Ono, Y. Nakamura et al., "Data mining methods for omics and knowledge of crude medicinal plants toward big data biology," *Computational and Structural Biotechnology Journal*, vol. 4, no. 5, pp. 1–14, 2013.

[43] D. C. Mohr, M. N. Burns, S. M. Schueller, G. Clarke, and M. Klinkman, "Behavioral intervention technologies: evidence review and recommendations for future research in mental health," *General Hospital Psychiatry*, vol. 35, no. 4, pp. 332–338, 2013.

[44] J. M. Ansermino, "From the journal archives: improving patient outcomes in the era of big data," *Canadian Journal of Anesthesia*, vol. 61, no. 10, pp. 959–962, 2014.

[45] T. Klingström, L. Soldatova, R. Stevens et al., "Workshop on laboratory protocol standards for the molecular methods database," *New Biotechnology*, vol. 30, no. 2, pp. 109–113, 2013.

[46] J. Mervis, "U.S. Science policy: agencies rally to tackle big data," *Science*, vol. 335, no. 6077, p. 22, 2012.

[47] Y. Mohammed, E. Mostovenko, A. A. Henneman, R. J. Marissen, A. M. Deelder, and M. Palmblad, "Cloud parallel processing of tandem mass spectrometry based proteomics data," *Journal of Proteome Research*, vol. 11, no. 10, pp. 5101–5108, 2012.

[48] J. Karlsson and O. Trelles, "MAPI: a software framework for distributed biomedical applications," *Journal of Biomedical Semantics*, vol. 4, no. 1, article 4, 2013.

[49] M. R. Bower, M. Stead, B. H. Brinkmann, K. Dufendach, and G. A. Worrell, "Metadata and annotations for multi-scale electrophysiological data," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, vol. 2009, pp. 2811–2814, 2009.

[50] S. Ranganathan, C. Schönbach, J. Kelso, B. Rost, S. Nathan, and T. W. Tan, "Towards big data science in the decade ahead from ten years of InCoB and the 1st ISCB-Asia Joint Conference," *BMC Bioinformatics*, vol. 12, supplement 13, p. S1, 2011.

[51] M. V. DiLeo, G. D. Strahan, M. den Bakker, and O. A. Hoekenga, "Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome," *PLoS ONE*, vol. 6, no. 10, Article ID e26683, 2011.

[52] C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng, "Big data bioinformatics," *Journal of Cellular Physiology*, vol. 229, no. 12, pp. 1896–1900, 2014.

[53] L. Dai, X. Gao, Y. Guo, J. Xiao, and Z. Zhang, "Bioinformatics clouds for big data manipulation.," *Biology direct*, vol. 7, article 43, 2012.

[54] D. MacLean and S. Kamoun, "Big data in small places," *Nature Biotechnology*, vol. 30, no. 1, pp. 33–34, 2012.

[55] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *The Journal of the American Medical Association*, vol. 309, no. 13, pp. 1351–1352, 2013.

[56] V. Marx, "Biology: the big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[57] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis," *Nature Reviews Genetics*, vol. 11, no. 9, pp. 647–657, 2010.

[58] J. B. Cole, S. Newman, F. Foertter, I. Aguilar, and M. Coffey, "Breeding and genetics symposium: really big data: processing and analysis of very large data sets," *Journal of Animal Science*, vol. 90, no. 3, pp. 723–733, 2012.

[59] "Finding correlations in big data," *Nature Biotechnology*, vol. 30, no. 4, pp. 334–335, 2012.

[60] E. Kolker, E. Stewart, and V. Ozdemir, "Opportunities and challenges for the life sciences community," *OMICS: A Journal of Integrative Biology*, vol. 16, no. 3, pp. 138–147, 2012.

[61] R. P. Troiano, J. J. McClain, R. J. Brychta, and K. Y. Chen, "Evolution of accelerometer methods for physical activity research," *British Journal of Sports Medicine*, vol. 48, pp. 1019–1023, 2014.

[62] E. Feldmann and D. S. Liebeskind, "Developing precision stroke imaging," *Frontiers in Neurology*, vol. 5, article 29, 2014.

[63] D. E. Green and E. J. Rapp, "Can big data lead us to big savings?" *Radiographics*, vol. 33, no. 3, pp. 859–860, 2013.

[64] B. A. Huberman, "Sociology of science: big data deserve a bigger audience," *Nature*, vol. 482, no. 7385, p. 308, 2012.

[65] C. Lynch, "Big data: how do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.

[66] S. E. White, "De-identification and the sharing of big data," *Journal of American Health Information Management Association*, vol. 84, no. 4, pp. 44–47, 2013.

[67] J. Chen, F. Qian, W. Yan, and B. Shen, "Translational biomedical informatics in the cloud: present and future," *BioMed Research International*, vol. 2013, Article ID 658925, 8 pages, 2013.

[68] S. Mavandadi, S. Dimitrov, S. Feng et al., "Crowd-sourced BioGames: managing the big data problem for next-generation lab-on-a-chip platforms," *Lab on a Chip*, vol. 12, no. 20, pp. 4102–4106, 2012.

[69] D. Riley and M. Mittelman, "Maps, 'big data,' and case reports," *Global Advances in Health and Medicine: Improving Healthcare Outcomes Worldwide*, vol. 1, no. 3, pp. 5–7, 2012.

[70] S. Hoffman and A. Podgurski, "Big bad data: law, public health, and biomedical databases," *Journal of Law, Medicine and Ethics*, vol. 41, no. 1, pp. 56–60, 2013.

[71] J. Cockfield, K. Su, and K. A. Robbins, "MOBBED: a computational data infrastructure for handling large collections of event-rich time series datasets in MATLAB," *Frontiers in Neuroinformatics*, vol. 7, article 20, 2013.

[72] S. F. Martin, H. Falkenberg, T. F. Dyrlund, G. A. Khoudoli, C. J. Mageean, and R. Linding, "PROTEINCHALLENGE: crowd sourcing in proteomics analysis and software development," *Journal of Proteomics*, vol. 88, pp. 41–46, 2013.

[73] D. B. Lindenmayer and G. E. Likens, "Analysis: don't do big-data science backwards," *Nature*, vol. 499, no. 7458, article 284, 2013.

[74] S. Toh and R. Platt, "Big data in epidemiology: too big to fail?" *Epidemiology*, vol. 24, no. 6, article 939, 2013.

[75] F. X. Castellanos, A. Di Martino, R. C. Craddock, A. D. Mehta, and M. P. Milham, "Clinical applications of the functional connectome," *NeuroImage*, vol. 80, pp. 527–540, 2013.

[76] J. Currie, "'Big data' versus 'Big brother': on the appropriate use of large-scale data collections in pediatrics," *Pediatrics*, vol. 131, supplement 2, pp. S127–S132, 2013.

[77] A. Docherty, "Big data—ethical perspectives," *Anaesthesia*, vol. 69, no. 4, pp. 390–391, 2014.

[78] B. Shen, A. E. Teschendorff, D. Zhi, and J. Xia, "Biomedical data integration, modeling, and simulation in the era of big data and translational medicine," *BioMed Research International*, vol. 2014, Article ID 731546, 1 page, 2014.

[79] N. H. Shah, "Translational bioinformatics embraces big data," *Yearbook of Medical Informatics*, vol. 7, no. 1, pp. 130–134, 2012.