Quality and Curation of Medical Images and Data

17

Peter M. A. van Ooijen

Recent years have shown an explosive growth in the use of artificial intelligence (AI) and deep learning (DL) not in the least for medical applications. These new technological developments have started a whole new discussion on how we can use the vast amount of available data in health care for processing by these computerized systems. However, especially the applications in health care demand a high level of (patient) data privacy and security. Furthermore, increasingly the requirement of getting appropriate consent from the patient, client, or participant is enforced [1] leading to additional challenges when collecting (retrospective) data. Another concern is that—although an abundance of data are acquired in health care—much of the health-related data are unstructured and not standardized. The actual ownership of medical data is also part of this discussion where different ownership rules can be involved with original, de-identified, anonymized, and processed data. Questions that arise from this are, for example, what data are still personal data for an individual patient or participant in a clinical trial and who actually owns the data that is produced by self-learning computer systems.

Once these issues and questions are solved and data can be collected and used, in many cases the big data are collected with a specific goal in mind in which the focus is on data quantity instead of data quality. This can hamper proper implementation and even lead to incorrect processing of the data or incorrect conclusions [1, 2]. In the era of machine learning and deep learning, the old adage of computer science that defines "garbage in, garbage out" gained renewed meaning and importance and the quality assessment and curation of the (imaging) data for AI and DL is said to take up to 80% of the data scientists' time [3, 4].

This chapter discusses the issue of data quality by looking at the process of curation of medical images and other related data and the different aspects that are involved in this when moving forward in the era of AI.

17.1 Introduction

When trying to answer questions about curation of medical images and data in the era of AI, one first has to answer the questions what the definition of artificial intelligence is. Different sources provide different answers to this question. Three often heard and read definitions are:

P. M. van Ooijen (✉)
University of Groningen, University Medical Center
Groningen, Groningen, The Netherlands
e-mail: p.m.a.van.ooijen@umcg.nl

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_17

1. Artificial intelligence is a computerized system that exhibits behavior that is commonly thought of as requiring intelligence.
2. Artificial intelligence is the science of making machines do things that would require intelligence if done by man.
3. AI is the science and engineering of making intelligent machines, especially intelligent computer programs.

In short, in machine or deep learning the algorithmic rules are no longer put into the system by a human observer, but the machine uses input data and known outcomes as training data to develop the algorithm. Therefore, data quality is a very important issue since the development of the algorithm is directly linked to the (quality of the) data collection used. Keep in mind that the results provided by such systems are always preliminary since every new bit of data entered into the learning system potentially alters the algorithm. Therefore, over time data also need to be of a constant high quality in order to avoid degradation of the algorithm because of newly arriving data and knowledge. This requires not only data collection quality but also a process of curation of collected data to increase the value and usability.

The University of Illinois' Graduate School of Library and Information Science defines data curation as "the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities *enable data discovery and retrieval, maintain its quality*, *add value*, and provide for *reuse over time*. This new field includes *authentication, archiving, management, preservation, retrieval*, and *representation*." [5, 6]. This data curation process is deemed a requirement to achieve an imaging biobank or data repository that is findable and reusable [7].

Current estimations suggest a doubling of the total amount of data in the world every 2–3 years [2]. Simultaneously, the percentage of the data collected digital instead of analogue increased dramatically in the past two decades. Although no fixed numbers over a long period of time are published, we can assume that similar increases in data have occurred in the past decades concerning medical imaging. In the nineties of the twentieth century, the digitalization of imaging commenced with the introduction of standardized data structure and communication with DICOM (Digital Imaging and Communication in Medicine) and the development of picture archiving and communication systems (PACS). These allowed a more convenient and standardized collection of the imaging data and also could guarantee the long-term storage and accessibility of the imaging data, provided a proper storage medium and migration strategy is employed [8]. The data increase itself was triggered by the ever-growing requirement for high-quality imaging data and mainly pushed forward by the developments in computer tomography (CT) and magnetic resonance imaging (MRI).

The increase in digital data collection also lowered the threshold to acquire data and thus allowed higher sampling frequency with more comprehensive data, thus further increasing the amount of data produced. These different factors have led to the collection of multi-TB PACS archives over the years with a variety of information per patient from different modalities, sequences, protocols, etc. Also, post-processed data obtained during the analysis and review from a variety of tools and workstations can be included in the patient data in the PACS as well as reports and other meta-information. When conducting retrospective data collection from such a PACS environment, the challenge is to include the relevant selection from the dataset acquired and generated that can be used for analysis and will lead to the required insight. What data to collect and at which frequency is still a human decision and thus prone to error, variation, and personal or institutional preferences. Because of this, the risk of collecting largely useless data collections is present. Often it is those types of collections of questionable quality that have to be used in artificial intelligence and deep learning.

Different machine learning and deep learning systems are developed both supervised and non-supervised and new networks are being published frequently. Selection of the proper environment or network is therefore also part of the challenge of deep learning, and this selection should be adapted to the properties of the data collection used to train and test the network. Regardless of the system selected, the availability of an appropriate training dataset is vital [2]. The above demonstrates that in DL the quality of the dataset used is vital in every step of the development.

17.2 Data Discovery and Retrieval

As stressed before, data selected to be used as training input for artificial intelligence environments have to comply with high quality standards. The data need to be correct, have proper and validated labels, be accurate, be still "up to current standards," etc. However, even if the data are of high quality, it also needs to be of sufficient size since applying AI to too small datasets will not render significant findings because of lack of power. Therefore, the right data collection(s) must be found and if needed combined to obtain a sufficiently large amount of unbiased data including all possible variations [4, 9–11].

The discovery and retrieval of (imaging) data in health care has a dimension on its own in that it is almost always personal health data from an individual. This hampers the discovery and retrieval of (imaging) data because multiple factors have to be considered when collecting health-related retrospective data from the electronic medical record (EMR) or picture archiving and communication system (PACS) or when acquiring prospective data through clinical trials or population studies. In many instances, the tools to mine these clinical systems in a structured, meaningful, and easy fashion are lacking but required for obtaining the datasets useful and adequate to perform AI [3].

And then again, when the correct cohorts are identified and the required approvals are obtained, the variability in the data collection can be enormous. First of all, medical imaging equipment is far from standardized and imaging data from different hospitals using equipment from different vendors or the same vendor but different equipment generation or protocol used can be incomparable in their image presentation and diagnostic quality, not only because of the fast development of new equipment but also because of the (subtle) differences in the technical implementation and scan sequences used by the different vendors.

Furthermore, these sequences for specific clinical questions are also not standardized and will provide different images based on the local preferences of a certain department or even a specific radiologist. Also, the variety in the naming of the protocols used by different vendors (especially in MRI) is decreasing the quality of the data. Therefore, the development of guidelines to data acquisition and standardization of protocols is a requirement to allow the construction of large and above all useful data collections [9]. Additionally, the imaging data that are usually collected nowadays are based on the already processed data in the shape of DICOM images while the raw data from which these human interpretable images originate (e.g., the k-space data of MR and sinograms of CT) are not stored while these could be a valuable source, with possibly less variation, for computerized analysis [3, 12].

Another major question to consider is the ownership and control of the data. The legal perspective concerning this question is covered in another chapter of this book, but there is also a more practical question to consider. Where does the data reside? In health care, we have observed a slow movement from hospital-centric data model to a more patient-centric data model. This also means integration of new information in this patient-centric model through for example the Internet of Things (IoT) and wearables. Furthermore, open science and open data are increasingly advocated by governments and funding agencies resulting in large collections of data mostly available in the cloud establishing sandbox environments to be used by anyone to train and validate their software [11].

The risk of putting data into the cloud is that we are in a sense losing control of this data, and thus discovery and retrieval of relevant data is severely hampered by the fact that we upload all our health-related information into a variety of dispersed non-connected and non-standardized cloud solutions [13].

The question arising from all this is, even if enough high-quality data are collected, are we able to find (or discover) the right data. And if we find certain data, are we able and allowed to actually retrieve the information contained in the system and combine different sources unambiguously into one single dataset. If we are able to gather the information contained in those different databases, it might bring us the capability of data merging and such obtain a new linked dataset with much richer information. However, this could also have implications on the usability of the data since combining multiple datasets could infringe the privacy of the individual that could not be recognized in the separate datasets but is identifiable by the combined data through data linkage.

The legal obligation to protect the privacy of the patient or participant is one of the crucial things to take care of when collecting data in health care [10, 11]. Current methodologies for anonymization and de-identification are often suboptimal [14]. Furthermore, the anonymization or de-identification has to be performed such that the scientific research value of the data is retained in the de-identified dataset while still removing all personal health information [15]. Therefore, new algorithms should be developed to conceal identities effectively both protecting the individual privacy and still maintaining the full value of the same data for analysis. These three aspects of de-identification, privacy, and data value can work against each other with opposing requirements and struggle with variability in data content and lack of standardization, thus hampering the automation of this process. Current repositories of research data such as the TCIA [7, 16] still have a workflow in place where curators visually check and when needed correct every DICOM file (image and header) entered into their database to ensure data privacy and correct handling of the data.

One specific challenge here is the fact that DICOM headers may contain proprietary information that is not part of the standard DICOM but could include information on the acquisition or nature of the imaging data enclosed that is vital for adequate advanced (post)processing of the data. However, these private tags may also include references to personal health information (PHI) or other information that could infringe the privacy of the subject [12]. The same holds for the comment fields that are available in the DICOM header, content of these fields is free text, and their use is often depending on local conventions. This content may thus vary per hospital or even per modality within a hospital. Therefore, these fields could also contain PHI manually entered by a technician or radiologist.

Besides the header information included with the DICOM file, the actual image contents may also pose the risk of disclosing privacy sensitive information. For example, in secondary captures, topograms, and ultrasound examinations where in each of these exams sensitive information can be burned into the image, removing this information is possible but difficult to automate since the location at which the sensitive information is stored in the images may vary and can be difficult to detect automatically. Furthermore, so-called DICOM containers can also be constructed where the DICOM header is present but instead of an image another file type is included into the file such as a PDF file. These files could even be full patient reports with all PHI included. Another special kind of DICOM file that needs to be handled with care is the DICOM Structured Report (SR). The SR file typically holds the report of the radiologist describing the image review and conclusion and thus could also reveal sensitive information depending on local policies or the reporting method of the individual radiologist.

A final challenge that needs to be identified is the fact that facial features can easily be obtained from MR and CT datasets of the head. By performing surface or volume rendering reconstruction of those datasets, the face of the subject involved becomes visible. Studies have shown that facial recognition technology is able to combine these reconstructions with pictures from, for example, social media profiles to reveal the identity of the imaged subject [17, 18]. Especially since name tagging of pictures in modernday social media directly links the person's name to the facial features.

The importance of careful curation of the data because of privacy risks has been reported on by different studies where the ability to breach the personal health information privacy was demonstrated on de-identified dataset. One example by Sweeney [19] shows that in 35 cases of an anonymized dataset obtained from a US hospital re-identification of the studies involved was possible by cross-linking to publicly available newspaper stories about hospital visits in the area.

17.3 Data Quality

The main reason for performing data curation is to increase the data quality of the data collection. However, an important consideration to start with is the question if the data quality is sufficient and useful for their application to artificial intelligence in the first place. It can be argued that when using big data an occasional bad sample or outlier will have little effect on the algorithm because of the large number of correct samples. However, the data richness also implies that the machine learning environment could use faulty inputs to determine the algorithm causing it to work on the training and testing dataset, but not in general use. One well-known example of AI and DL using suboptimal input datasets is a situation where the network is trained on a large multicenter database and with a test set performs adequate. However, at more careful inspection it is evident that the network is not trained to identify the pathology in the images but to recognize the features of a specific imaging device or hospital of origin because of unbalance with respect to the incidence of pathology in the dataset.

When looking at deep learning and machine learning as a system where the data together with the model are used to eventually come to a prediction, it is evident that the success of this system is not only depending on the quality of the model, but also on the quality of the data [20]. If the data, the model or both are of insufficient quality, the prediction will not be reliable.

The quality of the data used is thus essential for the validity of the outcome. A paper by Chalkidou et al. [21] showed that the current practice with data science and artificial intelligence leads to false discoveries because of fundamental flaws in the way the studies are performed. The issues that occur with those studies are a small sample size (12–72 cases, mean 44 cases) of often heterogenous cohorts, selection bias, and missing validation dataset (only 3/15 examined studies had a validation dataset).

There are multiple challenges defined that could negatively affect the quality of a dataset. These challenges are poor data collection practice, missing or incomplete values, non-standardized inconvenient storage solutions, intellectual property, security, and privacy [20, 22]. Assessing the quality of a dataset can therefore be challenging. To increase the use of data quality measures of datasets, multiple suggestions have been made to introduce some kind of data quality or maturity model. By assessing the dataset against such a model, the quality can be determined more objectively and possible use of the dataset is more evident.

One such a model for data quality was proposed by Lawrence [20]. He proposed to introduce a three-band model with subdivision into different levels per band. In this model, C4 would be the worst dataset and A1 the best. Band C would look at accessibility of the data. This could vary from C4 where the data might exist, but existence isn't even verified to C1 where data are collected in a standardized and known format and ready to be used without any constraints on the use. The next band, band B, would be about faithfulness and representation of the data. In this band, questions should be answered such as: Is the data that we got also what we expected? How are missing or incorrect values handled? What kind of encoding is used for the different data fields? How was the data collected? Is there bias in the dataset? Etc. In this band, the top quality would be B1 where we have a dataset that is C1 and where the limitations of the data are known to the user. Band A puts the data into the context. Here the ultimate question has to be answered if the dataset is appropriate to get to the correct prediction. It could be that in this phase expert annotation of the existing data or collection of additional data is required. Here level A1 would be curated data that are adjusted properly to allow getting the answer to the (clinical) question.

Based on the model by Lawrence, Harvey later introduced a version describing four data quality levels A–D more targeted to the medical domain [22]. The levels run from D where data are inaccessible, with unknown format and un-anonymized (current EMRs and PACSs in hospitals). In level C, anonymization is performed and ethical clearance obtained, but still the data are unstructured and show noise and gaps (EMR/PACS-based research collections). Level B introduces true representation with structured and visualizable data (structured and curated research collections). Finally, level A is a dataset containing contextual annotated and task ready data. According to Harvey, only A is AI usable data.

Although these kinds of models could be useful to categorize datasets for the purpose of machine learning, widespread application has not been established yet.

17.4 Adding Value

A report by EMC in 2014 [23] showed that in 2013 of the data collected in the global digital environment only 22% could be useful for analysis if—and only if—it would be properly tagged or characterized. However, they concluded that the tagging is mostly lacking in the collected data and that only 5% was valuable target-rich data. At that time they projected that in 2020 possible useful data would be increased to about 37% of all data collected with a doubling in target-rich data to about 10%.

In the case of medical imaging data, the proper annotation or tagging of the imaging data is also of vital importance [12] (level A of the model of Harvey described in the previous section). In order to train or validate AI and ML systems, a proper annotation is needed to define the ground truth that is used to learn and check results. However, no standardized syntax or method is available to collect the ground truth, and furthermore, the actual ground truth is difficult to obtain in most cases.

The two main standardized annotation methods are the Annotation and Image Markup (AIM) standard and the DICOM Presentation State (PS). AIM is developed within the National Cancer Informatics Program of the National Cancer Institute [10]. With AIM information is annotated and these annotations can be stored in a DICOM-compliant manner for later analysis. Although AIM is frequently reported to be used in research, it is not a widely accepted and used standard yet, and although DICOM PS is part of the globally accepted DICOM standard, it is still little used by software developers to report on annotations. Furthermore, clinically obtained annotations can in most cases not be used directly when performing AI because the annotations could contain personal health information which should not be present in research data. Therefore, annotations have to be redone when using the data for AI training and validation. Segmentation of the imaging data is even worse; no current widely accepted standard exists to store and communicate segmentation results between different tools from different vendors/sources.

The ground truth currently frequently used for training AI will result from a radiological report, a pathology examination report, or surgical reports. In this case, the value of the data on the "ground truth" relies both on the expertise of the observer describing the result, the accuracy of the description, and on the quality of the measurement methods. However, the accuracy of the results described by a physician is compromised by the fact that many reports are still free text without standardized lexicon or terminology resulting in multi-interpretable ambiguous reports with an abundance of synonyms. Furthermore, these different reports can even provide different measurements or conclusions and distinguishing which of these is the actual ground truth is a challenge that can often not be tackled. Natural language processing (NLP) could be a solution in situations where structured reporting and coding is not being used (as unfortunately still is the case in most hospitals).

17.5 Reuse Over Time

Part of the value of a dataset is the ability to use that dataset repeatedly over time. With the advent of bg data approaches, the data discovery and retrieval tend to shift from a targeted approach where specific data are collected to an approach where as much data as possible are gathered without a clear goal in mind because of possible future applications or novel insights that can be obtained [2]. When collecting data in this manner, assumptions have to be made on what data to collect and keep for future reference and use. Therefore, assessing the quality of this data collection is very cumbersome since the application of the data is still unknown. Furthermore, reuse also introduces other challenges and questions concerning the legal aspects of data privacy and intended use [2].

To allow reuse over time, the data should comply with the FAIR principle and be Findable, Accessible, Interoperable, and Reusable [24]. The FAIR guiding principles, that can be found in a table published by Wilkinson et al. at *(*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/*)*, define the principles that should be met to obtain a FAIR data collection.

To achieve this, a proper IT infrastructure is required to store these data [10] including an accurate description and indexing of the data. Such systems are often described in terms of and referred to as imaging biobanks. Imaging biobanks are defined as IT systems holding relevant data and allowing interoperability between them in a federated set-up [25]. Currently, multiple (research) institutes, scientific organizations, and funding agencies are advocating the opening up of imaging data for reuse over time and designing and building environments to allow this. Examples are the Cancer Imaging Archive (CIA), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), and the Osteoarthritis Initiative (OAI). Those archives or imaging biobanks contain collections of anonymized and curated (imaging) data that can be used for scientific purposes.

As an example, the CIA is a repository for cancer imaging and related information from the US National Cancer Institute [7]. With a content of over 30 million radiology images from over 37,500 subjects, it holds a wealth of information on cancer imaging. Data descriptions are used to categorize and organize the database into collections by tumor type. All data are manually curated and anonymized.

Although these initiatives exist, the need for better ways to construct FAIR data repositories is still prominently discussed, frequently also stressing the specific requirements to such datasets when they are to be used for machine and deep learning purposes [12].

17.6 Some Tools of the Trade

As in any application domain dealing with data, a vast number of tools are available to support in the different steps of the data curation process of medical data. There are tools for collection and anonymization of the data, for enrichment of the data, and for cleaning and curating the data. Without the illusion of being complete, Table 17.1 shows some examples of open source and freeware tools available for the different steps. When selecting tools to help you to obtain valid datasets, it is important that you select tools that are as simple as possible, and it might also help to restrict to using a defined set of tools within your research group or institution.

17.7 Conclusions

Only recently has data curation made the calendar of medical imaging research. Therefore, the understanding and role of data curation in the medical imaging domain is still limited. Often new research projects do not take into account the cost and manpower required to perform data curation either when collecting the data from the start (data curation "by design") or when data are collected from existing sources and, if needed, combined. However, in order to obtain datasets that can be used for future purposes, obtaining high-quality data is obligatory and data curation should be a requirement.

Table 17.1 List of examples of freely available tools for data handling and curation

Tool	Purpose	Where to find
CTP	Data collection/anonymization	https://www.rsna.org/ctp.aspx
TextAnonHelper	Text anonymization	https://bitbucket.org/ukda/ukds.tools.textanonhelper/wiki/Home
DeFacer	Anonymization by removal of facial features	https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface
POSDA	Archival and Curation of DICOM datasets	https://github.com/UAMS-DBMI/PosdaTools [26]
OpenRefine	Data cleaning tool	http://openrefine.org/
Colectica for Excel	Excel extension for documentation	https://www.colectica.com/software/colecticaforexcel/
Open Clinica	Clinical Data Management tool	https://www.openclinica.com/
RedCap	Clinical Data Management tool	https://www.project-redcap.org/
XNAT	Platform to support imaging-based research	https://www.xnat.org/

References

1. Rosenstein BS, et al. How will big data improve clinical and basic research in radiation therapy? Int J Radiat Oncol. 2015;95:895–904.
2. Mayer-Schonberger V, Ingelsson E. Big data and medicine: a big deal? J Intern Med. 2017.
3. Ridley EL. How to develop deep-learning algorithms for radiology. AuntMinnie.com. 2017. https://www.auntminnie.com/index.aspx?sec=sup&sub=aic&pag=dis&ItemID=118078. Accessed 6 June 2018.
4. Redman TC. If your data is bad, your machine learning tools are useless. Harv Bus Rev. 2018. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless. Accessed 6 June 2018.
5. U of Illinois. 2018. https://www.clir.org/initiatives-partnerships/data-curation/. Accessed 9 May 2018.
6. Freitas A, Curry E. Big data curation. In: Cavanillas JM, et al., editors. New horizons for a data-driven economy. Cham: Springer International Publishing; 2016.
7. Prior F, Smith K, Sharma A, Kirby J, Tarbox L, Clark K, Bennett W, Nolan T, Freymann J. Data descriptor: the public cancer radiology imaging collections of the Cancer Imaging Archive. Sci Data. 2017;4:170124.
8. van Ooijen PMA, Viddeleer AR, Meijer F, Oudkerk M. Accessibility of data backup on CD-R after 8 to 11 years. J Digit Imaging. 2010;23(1):95–9.
9. Aerts HJWL. Data science in radiology: a path forward. Clin Cancer Res. 2018;24(3):532–4.
10. Kansagra AP, Yu J-PJ, Chatterjee AR, Lenchik L, Chow DS, Prater AB, Yeh J, Doshi AM, Hawkins M, Heilbrun ME, Smith SE, Oselkin M, Gupta P, Ali S. Big data and the future of radiology informatics. Acad Radiol. 2016;23:30–42.
11. Tang A, Tam R, Cadrin-Chenevert A, Guest W, Chong J, Barfett J, Chepelev L, Cairns R, Mitchell R, Cicero MD, Gaudreau Poudrette M, Jaremko JL, Reinhold C, Gallix B, Gray B, Geis R. Canadian Association of Radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J. 2018;69:120–35.
12. Kohli M, Summers R, Geis R. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. J Digit Imaging. 2017;30:392–9.
13. Lupton D. Who owns your personal health and medical data? This Sociological Life BLOG. 2015.
14. Aryanto KYE, Oudkerk M, van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol. 2015;25(12):3685–95. https://doi.org/10.1007/s00330-015-3794-0.
15. Moore SM, et al. De-identification of medical images with retention of scientific research value. Radiographics. 2015;35:727–35.
16. Clark K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–57.
17. Prior FW, Brunsden B, Hildebolt C, et al. Facial recognition from volume rendered magnetic resonance imaging data. IEEE Trans Inf Technol Biomed. 2009;13(1):5–9.
18. Mazura JC, Juluru K, Chen JJ, Morgan TA, John M, Siegel EL. Facial recognition software success rate for the identification of 3D surface reconstructed facial images: implications for patient privacy and security. J Digit Imaging. 2012;25(3):347–51.
19. Sweeney L. Only you, your doctor, and many others may know. Technology Science. 2015. http://techscience.org/a/2015092903. Accessed 6 June 2018.
20. Lawrence ND. Data readiness levels. 2017. arXiv:1705.02245v1 [cs.DB].
21. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One. 2015;10:e0124165.

22. Harvey H. Is medical imaging data ready for Artificial Intelligence? AuntMinnieEurope. 2017. https://www.auntminnieeurope.com/index.aspx?sec=sup&sub=pac&pag=dis&ItemID=615032. Accessed 6 June 2018.
23. EMC. The digital universe of opportunities: rich data and the increasing value of the internet of things. Executive summary data growth, business opportunities, and the IT imperatives. EMC. 2014. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Accessed 9 June 2018.
24. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
25. ESR. ESR position paper on imaging biobanks. Insights Imaging. 2015;6(4):403–10.
26. Bennett W, Metthews J, Bosch W. SU-GG-T-262: open-source tool for assessing variability in DICOM data. Med Phys. 2010;37:3245.