Hanyang Medical Reviews

Hanyang Med Rev 2017;37:86-92
https://doi.org/10.7599/hmr.2017.37.2.86
pISSN 1738-429X eISSN 2234-4446

Review

Status and Direction of Healthcare Data in Korea for Artificial Intelligence

Yu Rang Park[1,2], Soo-Yong Shin[3]

[1]Department of Biomedical Informatics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Clinical Research Center, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[3]Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do, Korea

Recent rapid advances in artificial intelligence (AI), especially in deep learning methods, have produced meaningful results in many areas. However, to achieve meaningful results for healthcare through AI, it is important to understand the meaning and characteristics of data in that area. For medical AI, a simple approach that accumulates massive amounts of data based on existing big data concepts cannot provide meaningful results in the healthcare field. We need well-curated data as opposed to a simple aggregation of data. The purpose of this study is to present the types and characteristics of healthcare data and future directions for the successful combination of AI and medical care.

Key words: AI; Machine Learning; Healthcare Data; Smart Data

Corresponding Author: Soo-Yong Shin
Department of Computer Science and Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Korea
Tel: +82-31-201-2543
E-mail: sooyong.shin@khu.ac.kr

Received 1 Aug 2017
Revised 23 Oct 2017
Accepted 5 Nov 2017

INTRODUCTION

Recently, artificial intelligence (AI) has been highlighted in various areas including healthcare [1–4]. AI can be categorized into symbolic AI such as expert systems and machine learning (ML), which includes deep learning. Technically, recently mentioned AI refers to ML or deep learning. Deep learning, which is inspired by biological neurons, is a subcategory of machine learning algorithms [5]. Machine learning (including deep learning) requires a large amount of training data to improve performance. Therefore, to implement a good healthcare AI system, we need a vast amount of healthcare data. Many people believe there is a large amount of data in hospitals based on the wide adaptation of electronic medical records (EMR). They mentioned that the adoption rate of EMR in the United States was dramatically increased to 97% after the introduction of the Health Information Technology for Economic and Clinical Health (HITECH) Act [6] and the adoption rate of EMR in Korea is more than 92%. Nearly all hospitals in Korea also use the computerized physician order entry (CPOE) system. However, the EMR adoption rate is only 58.1%, and the fully comprehensive EMR adoption rate has dropped to 11.6% [7]. This implies a lack of digitalized data for healthcare AI research in Korea.

Even though there is a large amount of data, having only a large quantity of data based on big data concepts may fail to achieve an applicable healthcare AI system. We need well-curated and labeled data. For example, 54 US licensed ophthalmologists and ophthalmology senior residents have reviewed 128,175 retinal images to build a well-curated dataset [3]. Current digitalized medical records require more in-depth curation to be used for research. Moreover, to realize precision medicine with the aid of AI methods, we need many new healthcare data types including genome and wearable data.

Table 1. Types of healthcare data in South Korea

Type	Description	Main sources
Clinical data	Data collected during the course of ongoing patient care	Hospital information systems (EMR, CPOE, PACS, LIMS)
Claim data	Data generated by billing process	Public or private insurance providers
Research data	Published biomedical research data or clinical trial data	Pharmaceutical companies, regulators, international clinical trial repositories, biomedical journals
Genomic data	Human genome-related data	Research institutes, public databases, Hospital Information Systems by National conditional insurance for screening of NGS-based gene panel
Patient-generated health data	Health-related data created, recorded, or gathered by or from patients (or family members or other caregivers)	Smartphone apps, social media, wearable devices
Social determinants of health	The conditions where people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life	Governments, researchers

In this paper, we first categorize the diverse types of healthcare data that can be used for AI. Then, we analyze the status of healthcare data in Korea and propose a future direction for healthcare AI development.

TYPES OF HEALTHCARE DATA

The healthcare data of an individual can be collected from diverse sources [8]. Though there could be different categorization, we categorized healthcare data into six categories such as clinical data, claim data, research data, genomic data, patient-generated health data, and social determinants of health as shown in Table 1.

Clinical data are obtained at the point of care of a medical facility, hospital, clinic, or practice (within a clinical setting). Clinical data include demographic information, diagnosis, treatment, prescription drugs, laboratory tests, physiologic monitoring data, hospitalization, etc. Such data are regarded as the most important type of data in healthcare [9], if the clinical data can be stored in electronic formats or written in plain text. The representative information systems for clinical data are EMR, CPOE, the picture archiving and communication system (PACS), and the laboratory information management system (LIMS).

Claim data describe the billing information for insurance claims. Claim data can be accessed by a government agency (e.g., health insurance review and assessment service in Korea) or private insurance companies. The merit of claim data is that it can offer data on a large number of patients from multiple hospitals or clinics as well as the longitudinal data of a single person by combining claims.

Research data are the health-related data from the experimental results in biology laboratories, published research articles, and clinical trial data. This data can provide the most recent healthcare information. Though pharmaceutical companies are the major data holders, there are many public biomedical databases [10, 11].

Genomic data can be obtained from the study of genomes in an academy or from genomic/genetic tests in hospitals. Recently, the South Korean government has provided conditional insurance for the next-generation sequencing (NGS) technology-based cancer gene panel. Therefore, it can be included in research data such as biological study data or clinical data including cancer panel and genetic test data. Based on the rapid advances in NGS technologies, genomic data have been highlighted as the crucial data for personalized or precision medicine.

Patient-generated health data (PGHD) is health-related data created, recorded, or gathered by a patient [12]. In other words, PGHD are the health data that are collected outside of a clinical setting. Usually, PGHD are collected by healthcare wearables, home health monitoring devices, or self-reported methods.

Social determinants of health (SDOH) comprise the data of conditions in which people are born, grow, work, and live. In other words, SDOH are the wider set of forces and systems shaping the conditions of daily life, e.g., gender, social and political situations, weather, or environmental factors [13]. The focus of PGHD is on person-generated lifelog data, while

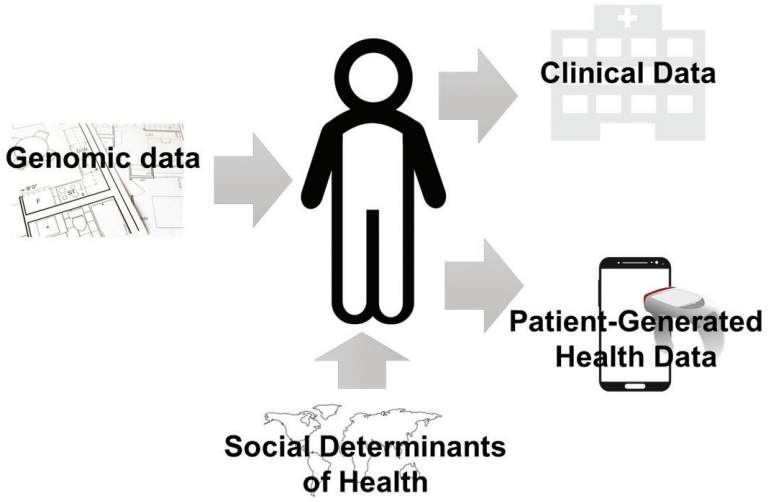

Fig. 1. Healthcare data and human health
Genomic data could be regarded as a blueprint, and SDOH can change the blueprint as we live. These two types of health data can be the input of a health condition. Clinical data and PGHD are the outcomes of a health condition.

the focus of SDOH is on environmental data that cannot be controlled by individuals.

In summary, as shown in Fig. 1, genomic data might be a blueprint of a health condition. SDOH are important factors in affecting health conditions as they can even change the blueprint, i.e. epigenomics. Clinical data and PGHD are the intermittent and continuous monitored outcomes of a health condition, respectively. Genomic data and SDOH are the inputs of a health condition, whereas, clinical data and PGHD are the outputs of a health condition. Research data can be both an input and output.

STATUS OF HEALTHCARE DATA AND FUTURE DIRECTION FOR WELL-CURATED DATA

1.Clinical data

To apply machine learning methods to healthcare data, the data should first be stored in an information system instead of physical documents. The important point is how well-organized and well-structured the data exist in the system, not how many data stores in the system. Though EMR was used, most of the data in EMR are unfortunately written in an unstructured text format. Physicians describe patient conditions using free text with many abbreviations. This implies that current EMR can be simply regarded as word processor files. Because the main purpose of the hospital information system 10 years ago in Korea was the implementation of four less hospital drudgery (paperless, chartless, slipless, and filmless), the system focused on only digitizing clinical data from handwritten papers. After a decade, with the arrival of the big data era, we face the limitation of free text data in current EMR, i.e., inconsistent meanings for equivalent abbreviations, and the incompleteness of data [14].

To overcome this limitation, first, we must implement the structured clinical template for accurate and complete data entry. There are diverse approaches including the clinical contents model (CCM) [15], International Organization for Standardization (ISO) 13606 [16, 17], detailed clinical model (DCM) [18], and Clinical Information Modeling Initiative (CIMI) [19]. By the implementation of the structured template, clinical data in EMR can be standardized to resolve ambiguity, provide semantic interoperability, and prevent data entry errors. Second, to extract the meaningful information from the existing text documents, clinical natural language processing (NLP) methods should be developed [20–22], or the simple regular expression can also be applied [23]. In Korea, the simple regular expression can be more practical and promising at the current stage due to the lack of Korean NLP research.

The remaining clinical data including order, lab, and image data are relatively well-structured when compared to EMR data because order data are coded data for billing purposes, the data in the LIMS usually contains numbers, and image data use the Digital Imaging and Communications in Medicine (DICOM)

standard. Currently, deep learning techniques are intensively applied to the medical image data in the PACS [3, 4], because deep learning methods have demonstrated their capabilities for image analysis in other areas. In addition, there are standards for image data, e.g., DICOM or JPEG. For the same reason, digital pathology images are good target areas for deep learning [24, 25]. If we have standardized clinical data equivalent to MRI and CT images, we can apply diverse machine learning technologies as well.

2. Claim data

After large data has been highlighted, many research articles using claim data have been published. Claim data in Korea can offer the data of nearly the entire Korean population owing to the national insurance services (97.1% of the total population). Therefore, the easiest way to increase the number of patients is to use the claim data. In addition, a single hospital does not have the lifelong medical history of a patient. Though claim data do not include outcomes of clinical practice, it can offer the longitudinal medical history reported by multiple hospitals. Currently, claim data and clinical data are the most widely used healthcare data types [9].

However, because claim data only contain diagnosis, medication, and laboratory test order information, a detailed analysis requiring laboratory test results cannot be performed. Another significant issue in claim data is the inconsistency between the clinical data from hospitals and claim data [26]. Claim data are generated by the billing staff in hospitals based on the diagnoses in the hospital information system. In this case, a degradation of the diagnosis-coding accuracy might be unavoidable. Therefore, claim data should be used carefully.

Interestingly, claim data can be used to improve the accuracy and completeness of clinical data. Because billing code is tightly coupled with the order code in the CPOE system, the order data is well coded and reviewed. This means that if we use standard terminology for claim data, every hospital will adopt the standard terminology used in billing. Therefore, the out-of-date billing code and system used in the Korean Health Insurance Review and Assessment Service should be changed to use up-to-date technologies such as HL7 FHIR (Fast Healthcare Interoperability Resources), which will be used for the US Precision Medicine Initiative [27].

3. Research data

The importance of research data has increased owing to the rapid advances in human genome research. Though the clinical guidelines are updated periodically, there is some delay in the update. Therefore, to apply treatment based on genomic data or to treat patients with rare diseases, research data should be considered. However, this data should be used with caution because research results are revised continuously. Moreover, to utilize research data, an additional IT system is necessary to connect the databases [10, 11] that are located outside of the hospitals.

4. Genomic data

Genomic data are the most highlighted data recently in the healthcare domain. When performing clinical sequencing, a vast amount of data with a complicated series of processes is generated [28, 29]. These complicated series of processes including sequencing and analytic pipelines imply several limitations on genomic data including data quality and reliability. To be used in clinical practice, clinical sequencing results should provide the same interpretation regardless of the performing laboratories or hospitals. However, because of the different sequencing platforms and analysis pipelines, the reliability of clinical sequencing is in doubt. In addition, even though the variants are equivalent, clinical interpretation results can be different. There is no gold standard knowledgebase for genomic data interpretation. Therefore, the efforts in quality control [30] and genotype–phenotype knowledgebase construction [31] should be followed.

Moreover, there is another problem in the integration of genomic data with clinical data for clinical practice. To overcome this problem, several international standards have been developed in the ISO/TC (Technical Committee) 215 Health informatics [32] and HL7 [27].

5. Patient-generated health data

PGHD is promising data for healthcare in terms of data size and clinical relevance because healthcare Internet of Things (IoT) devices or wearables can support the continuous monitoring of each person in everyday life. As shown in Fig. 2, two of the most promising data types in the next 5 years are genomic data and PGHD. PGHD can complement clinical data, which are intermittent data, because PGHD are continuously collected. However, PGHD are not well incorporated into the current clinical practice because they are emerging fields, and their characteristics are different from those of traditional clinical data. Recently, the US Food and Drug Administration (FDA) has started to pay attention to PGHD [33, 34]. The FDA

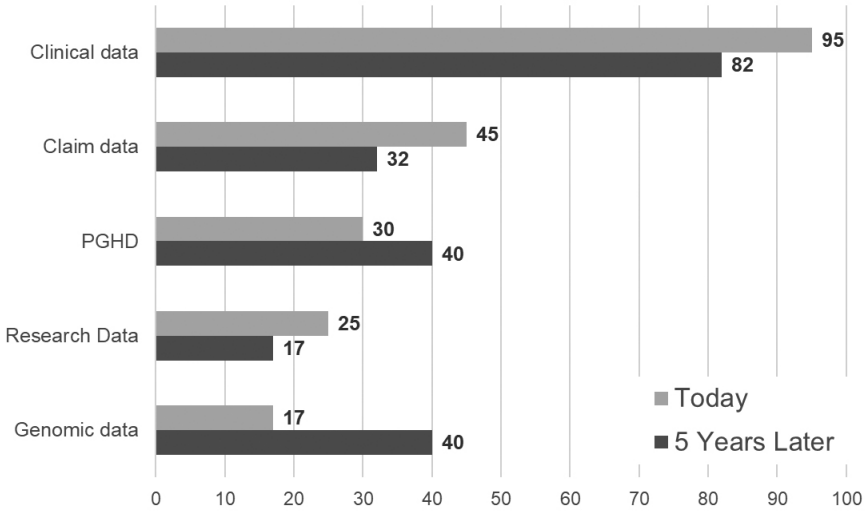

Fig. 2. Most useful sources of healthcare data today and in 5 years.
Modified from the figure on page 1 of [9] by selecting the healthcare data types matched in this paper.

uses terms such as real-world data (RWD) and real-world evidence (RWE). RWD are collected from sources outside of traditional clinical trials, contained within medical devices, and in tracking patient experience during care, including in-home use settings [33]. RWE is derived from the aggregation and analysis of RWD elements [33].

The limitation of PGHD, especially the PGHD collected from wearables, is its accuracy and interoperability. As technologies are developed, its accuracy will be improved to match the accuracy of medical devices, or a different category of healthcare wearables can be considered for the characteristics of continuous monitoring as opposed to the intermittent data of medical devices. To improve the interoperability, several standard organizations such as the Open Connectivity Foundation (OCF) [35] or International Electrotechnical Commission (IEC)/TC 124 Wearable Electronic Devices and Technologies [36] have started to implement standards.

6. Social determinants of health

SDOH can be regarded as epidemiology data. This means SDOH have a long history of patient treatment and are proven as an important factor for health outcome. For example, the zip code of a person can affect the healthcare conditions more than genomic factors. Therefore, SDOH should be actively included in clinical practice [37] and AI development.

AI researcher can easily utilize PGHD and SDOH compared to clinical data since those data can be directly collected from the individual without help of the hospitals. Therefore, new ideas for wellness care with AI can be promising.

SUMMARY

Machine learning technologies have demonstrated their use (or feasibility) in the healthcare domain. However, current healthcare AI research does not fully utilize diverse healthcare data. Healthcare data generators including physicians should try to improve the reliability and accuracy of healthcare data. We must keep in mind that the data used is more important than an algorithm [38]. As in Datasets Over Algorithms[38], the average number of years for a breakthrough to occur is 3 for datasets and 18 for algorithms. Breakthroughs can be achieved using data sets and not algorithms.

To improve the quality of healthcare data and collect multi-institutional data, healthcare data standards should be adopted in hospitals. Recently, the common data model (CDM) has been highlighted to perform big data research as the method of collecting multi-institutional research. OHDSI (Observational Health Data Science and Informatics) OMOP (Observational Medical Outcomes Partnership) CDM [39] is the most popular CDM in Korea. However, there are several other CDMs such as PCORnet [40] and Sentinel [41]. Researchers can choose the necessary CDM based on their research purpose. The important concept of a CDM is that the data interoperability is guaranteed based on standard terminologies. Also, the CDM

can accommodate different types of observational healthcare data in one standardized, easy-to-use format. Because of this, different types of observational analyses based on AI can be implemented using the same basic building blocks.

Finally, healthcare AI cannot escape from ethical issues such as privacy. The regulations to protect patient privacy may create legal barriers to the flow of patient data to applications. This barrier could be considered as an impediment to developing AI methods. However, for the long-term and rational development of healthcare AI, we always try to protect the privacy of patients. To protect privacy, de-identification methods for healthcare data should be developed. In addition, recent privacy preserving data mining methods including differential privacy or homomorphic encryption should be applied to healthcare data [42].

REFERENCES

1. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform 2017; 21(1): 4–21.
2. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016; 6: 26094.
3. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016; 316(22): 2402.
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639): 115–118.
5. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press:2017:168-177.
6. Gold M, McLaughlin C. Assessing HITECH Implementation and Lessons: 5 Years Later. Milbank Q 2016; 94(3): 654–687.
7. Kim YG, Jung K, Park YT, Shin D, Cho SY, Yoon D, et al. Rate of electronic health record adoption in South Korea: A nation-wide survey. Int J Med Inform 2017; 101: 100–107.
8. Weber G, Mandl K, Kohane I. Finding the Missing Link for Big Biomedical Data. JAMA 2014; 2115: 1–2.
9. Compton-Phillips A. Care Redesign Survey: What Data Can Really Do for Health Care [Internet]. New England Journal of Medicine 2017. [cited 2017 July 13]. Available from: http://catalyst.nejm.org/effectiveness-healthcare-data-survey-analysis/.
10. National Center for Biotechnology Information. All Resources - Site Guide - NCBI [Internet]. [cited 2017 July 8]. Available from: https://www.ncbi.nlm.nih.gov/guide/all/.
11. European Bioinformatics Institute. Bioinformatics services [Internet]. [cited 2017 July 8]. Available from: https://www.ebi.ac.uk/services.
12. HealthIT.gov. Patient-Generated Health Data [Internet]. [cited 2017 July 8]. Available from: https://www.healthit.gov/policy-researchers-implementers/patient-generated-health-data.
13. World Health Organization. Social Determinants of Health [Internet]. [cited 2017 July 8]. Available from: http://www.who.int/social_determinants/en/.
14. Kho AN, Pacheco J a, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. Sci Transl Med 2011; 3(79): 79re1.
15. Yun JH, Ahn SJ, Kim Y. Development of clinical contents model markup language for electronic health records. Healthc Inform Res 2012; 18(3): 171–177.
16. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT. Towards ISO 13606 and openEHR archetype-based semantic interoperability. Stud Health Technol Inform 2009; 150: 260–264.
17. The CEN/ISO EN13606 standard [Internet]. [cited 2017 July 7] Available from: http://www.en13606.org/the-ceniso-en13606-standard 2017].
18. Goossen W, Goossen-Baremans A, van der Zel M. Detailed Clinical Models: A Review. Healthc Inform Res 2010; 16(4): 201.
19. OpenCIMI. Mission and Goals | www.opencimi.org [Internet]. [cited 2017 July 7]. Available from: https://www.opencimi.org/.
20. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiol 2016; 279(2): 329–343.
21. Demner-Fushman D, Elhadad N. Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. Yearb Med Inform 2016(1): 224–233.
22. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2015: Leveraging the Variety of Texts of Clinical Interest. IMIA Yearbook 2016(1): 234–239.
23. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A De-identification method for bilingual clinical texts of various note types. J Korean Med Sci 2015; 30(1): 7–15.
24. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NNC, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Sci Rep 2017; 7: 46450.
25. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal 2016; 33: 170–175.
26. Society for Participatory Medicine. Imagine someone had been managing

your data, and then you looked [Internet]. [cited 2017 July 13] Available from: https://participatorymedicine.org/epatients/2009/04/imagine-if-someone-had-been-managing-your-data-and-then-you-looked.html.
27. HL7. Index - FHIR v3.0.1 [Internet]. [cited 2017 July 13]. Available from: https://www.hl7.org/fhir/.
28. Evans JP, Powell BC, Berg JS, T M, BK R. Finding the Rare Pathogenic Variants in a Human Genome. JAMA 2017; 317(18): 1904.
29. Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL. Organizing knowledge to enable personalization of medicine in cancer. Genome Biol 2014; 15(8): 438.
30. US FDA. PrecisionFDA [Internet]. [cited 2017 July 13]. Available from: https://precision.fda.gov/.
31. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov. Forthcoming 2017. DOI: 10.1158/2159-8290.CD-17-0151.
32. International Organization for Standardization. ISO/TS 20428:2017 - Health informatics -- Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records [Internet]. [cited 2017 July 8]. Available from:https://www.iso.org/standard/67981.html.
33. US FDA. Use of Real-World Evidence to Support Regulatory Decision-Making [Internet]. [cited 2017 July 18]. Available from: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm513027.pdf
34. Harvard Business School. Evaluating New Drugs with Wearable Technology – Technology and Operations Management [Internet]. [cited 2017 July 8]. Available from: https://rctom.hbs.org/submission/evaluating-new-drugs-with-wearable-technology/.
35. OPEN CONNECTIVITY FOUNDATION (OCF) [Internet]. [cited 2017 July 11]. Available from: https://openconnectivity.org/.
36. International Electrotechnical Commission. IEC - TC 124 [Internet]. [cited 2017 July 11]. Available from: http://www.iec.ch/dyn/www/f?p=103:7:0::::FSP_ORG_ID,FSP_LANG_ID:20537,25.
37. Andermann A, CLEAR Collaboration. Taking action on the social determinants of health in clinical practice: a framework for health professionals. CMAJ 2016; 188(17–18): E474–E483.
38. KDDNugget. Datasets Over Algorithms [Internet]. [cited 2017 July 11]. Available from: http://www.kdnuggets.com/2016/05/datasets-over-algorithms.html.
39. Observational Health Data Sciences and Informatics. Common Data Model Documentation [Internet]. [cited 2017 July 11]. Available from: [http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm#common_data_model.
40. The National Patient-Centered Clinical Research Network. PCORnet Common Data Model (CDM) - PCORnet [Internet]. [cited 2017 July 11]. Available from: http://www.pcornet.org/pcornet-common-data-model/.
41. Sentinel. Distributed Database and Common Data Model | Sentinel System [Internet]. [cited 2017 July 11]. Available from: https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model.
42. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014; 15(6): 409–421.